ABSTRACT

Title of dissertation: COMPUTING WITH TRAJECTORIES: CHARACTERIZING DYNAMICS AND CONNECTIVITY IN SPATIOTEMPORAL NEUROIMAGING DATA

Manasij Venkatesh

Dissertation directed by: Professor Luiz Pessoa,
Department of Psychology
Professor Joseph F. JaJa,
Department of Electrical and Computer Engineering

Human functional Magnetic Resonance Imaging (fMRI) data are acquired while participants engage in diverse perceptual, motor, cognitive, and emotional tasks. Although data are acquired temporally, they are most often treated in a quasi-static manner. Yet, a fuller understanding of the mechanisms that support mental functions necessitates the characterization of dynamic properties. Firstly, we describe an approach employing a class of recurrent neural networks called reservoir computing, and show their feasibility and potential for the analysis of temporal properties of brain data. We show that reservoirs can be used effectively both for condition classification and for characterizing lower-dimensional "trajectories" of temporal data. Classification accuracy was approximately 90% for short clips of "social interactions" and around 70% for clips extracted from movie segments. Data representations with 12 or fewer dimensions (from an original space with over 300) attained classification accuracy within 5% of the full data. We hypothesize that such

low-dimensional trajectories may provide "signatures" that can be associated with tasks and/or mental states. The approach was applied across participants (that is, training in one set of participants, and testing in a separate group), showing that representations generalized well to unseen participants.

In the second part, we use fully-trained recurrent neural networks to capture and characterize spatiotemporal properties of brain events. We propose an architecture based on long short-term memory (LSTM) networks to uncover distributed spatiotemporal signatures during dynamic experimental conditions. We demonstrate the potential of the approach using naturalistic movie-watching fMRI data. We show that movie clips result in complex but distinct spatiotemporal patterns in brain data that can be classified using LSTMs ($\approx 90\%$ for 15-way classification), demonstrating that learned representations generalized to unseen participants. LSTMs were also superior to existing methods in predicting behavior and personality traits of individuals. We propose a dimensionality reduction approach that uncovers low-dimensional trajectories and captures essential informational properties of brain dynamics. Finally, we employed saliency maps to characterize spatiotemporally-varying brain-region importance. The spatiotemporal saliency maps revealed dynamic but consistent changes in fMRI activation data. Taken together, we believe the above approaches provide a powerful framework for visualizing, analyzing, and discovering dynamic spatially distributed brain representations during naturalistic conditions.

Finally, we address the problem of comparing functional connectivity matrices obtained from temporal fMRI data. Understanding the correlation structure

associated with multiple brain measurements informs about potential "functional groupings" and network organization. The correlation structure can be conveniently captured in a matrix format that summarizes the relationships among a set of brain measurements involving two regions, for example. Such functional connectivity matrix is an important component of many types of investigation focusing on network-level properties of the brain, including clustering brain states, characterizing dynamic functional states, performing participant identification (so-called "fingerprinting"), understanding how tasks reconfigure brain networks, and inter-subject correlation analysis. In these investigations, some notion of proximity or similarity of functional connectivity matrices is employed, such as their Euclidean distance or Pearson correlation (by correlating the matrix entries). We propose the use of a geodesic distance metric that reflects the underlying non-Euclidean geometry of functional correlation matrices. The approach is evaluated in the context of participant identification (fingerprinting): given a participant's functional connectivity matrix based on resting-state or task data, how effectively can the participant be identified? Using geodesic distance, identification accuracy was over 95% on resting-state data, and exceeded the Pearson correlation approach by 20%. For whole-cortex regions, accuracy improved on a range of tasks by between 2% and as much as 20%. We also investigated identification using pairs of subnetworks (say, dorsal attention plus default mode), and particular combinations improved accuracy over whole-cortex participant identification by over 10%. The geodesic distance also outperformed Pearson correlation when the former employed a fourth of the data as the latter. Finally, we suggest that low-dimensional distance visualizations based

on the geodesic approach help uncover the geometry of task functional connectivity in relation to that during resting-state. We propose that the use of the geodesic distance is an effective way to compare the correlation structure of the brain across a broad range of studies.

Computing with Trajectories: Characterizing Dynamics and Connectivity in Spatiotemporal Neuroimaging Data

by

Manasij Venkatesh

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2020

Advisory Committee:
Professor Luiz Pessoa, Chair/Advisor
Professor Joseph F. JaJa, Advisor
Professor Jonathan Z. Simon
Professor Behtash Babadi
Professor Daniel A. Butts, Dean's Representative

To Amma, Appa, and Akshay.

Acknowledgments

First and foremost, I would like to thank my family. Despite having a modest educational background, my parents made unfathomable sacrifices to ensure that my brother and I were never found wanting. This dissertation is a culmination of their vision from the first moments of my life. My brother, Akshay, is the first Ph.D. graduate in our family. The impact he has had on me cannot be overstated. Each step I took was made easier because he had already navigated the path before me. I'm grateful for his love, support, and birthday gifts each year.

I was fortunate to have two great advisors, Luiz Pessoa, and Joseph JaJa. Luiz was the single biggest influence on me in graduate school. His extraordinary energy never ceased to amaze me. He challenged me to tackle difficult problems and think critically about my own work. Several key ideas in this dissertation started with him walking into the lab for a casual conversation, often ending in vivid pictures on the whiteboard. I have truly benefited from all my conversations with him, and I'm sure I will carry forward so many of the things he taught me.

From my first interactions with Joseph, it was evident that his support was unwavering. Apart from making sure that I had all the computing resources I need, lab spaces to work, and the freedom to choose my own problems, he taught me to be patient and reflect on my own work. His gentle nudges were the primary reason

for the timely completion of this dissertation. I will never forget the kindness and patience with which he approached every situation. I hope it has rubbed off on me.

I'm also grateful to my committee members Jonathan Simon, Behtash Babadi, and Daniel Butts. Behtash was my first instructor at graduate school. His course on random processes inspired me and reminded me of how much I had missed school. Several of Jonathan's suggestions during the proposal of this thesis formed an important part of the later work. I thank Dan for his infectious enthusiasm and for encouraging me to think of the broader implications of my work. It was an honor to present my work to these members.

Several contributions to a Ph.D. don't get mentioned in papers and journals. But each lab member in Luiz's and Joseph's lab contributed to this dissertation in so many ways. Conversations with them helped me keep up with the latest research and gave me ideas to use them in my own work. I'm extremely grateful to Emily and Melanie, the program managers in ECE. Their support at the beginning of my Ph.D. through teaching assistantships came when it was much needed. I would also like to thank my undergraduate mentor, Dr. Deepu Vijayasenan, for introducing me to machine learning and guiding me to graduate school.

Five years ago, I moved to the United States without knowing anyone in Maryland. I'm grateful for all the friendships I made here. I had some of the most interesting, thought-provoking, and hilarious conversations (often occurring together) with Karthik and Nidhi. I could always count on them to be a knock away at Iribe. Racing for the UMD cycling team resulted in scars and broken bones but also in some of my strongest friendships. I will always cherish the long rides

through Montgomery Country and Annapolis with Will and Alex. Many thanks to Andy Huang and Joe Barrow, for making 8805 37th Ave such a fun place to come home to. I will miss the banter and all the pizza. Finally, I would like to thank my partner, Tunga, for her constant encouragement, love, and support. She took care of everything for me so I could focus on writing this dissertation. And perhaps most importantly, she taught me how to enjoy and celebrate the end.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Humans engage in diverse perceptual, motor, cognitive, and emotional tasks. For over a century, neuroscience researchers have investigated how the human brain receives multiple inputs, integrates new information with the past, and performs functions to generate behavior. Efforts to understand brain function based on lesion studies date back to the second half of the 1800s. For example, Broca, in one of the most important works on brain function localization, observed and concluded that a lesion of the left frontal lobe resulted in a loss of speech [1].

Recent technological developments have allowed for the acquisition of spatiotemporal neural data through invasive modalities such as electrocorticography (ECoG) and calcium imaging, and non-invasive modalities such as Magneto/Electroencephalography (M/EEG) and functional Magnetic Resonance Imaging (fMRI). They have provided further insights into the mapping from brain structure to function and suggest that brain regions are richly connected and participate in diverse functions [2].

Despite several advances, most experiments in fMRI are tightly controlled and often analyzed using univariate approaches that explain how various regions are linked to behavior. Whereas such experiments have provided valuable insights

into how the brain functions, more "naturalistic" stimuli likely engage the brain in other ways. The brain is a complex dynamic system; its ability to assimilate and process spatial and temporal features of stimuli is indispensable for naturalistic tasks. Given the rich spatially-and-temporally varying nature of neuronal responses, brain dynamics must be addressed head-on. Yet, a general computational framework for processing such data remains elusive [3].

Despite the potential of fMRI to be used to investigate temporal properties of brain data, most techniques are employed in a largely "static" fashion. That is, inputs to models are patterns of activation that are averaged across time [4]. Some studies have represented temporal information using an additional spatial dimension by considering a temporal data segment instead of the average signal during that period [5, 6]. Despite some progress, how temporal information is integrated across time, and questions regarding the dimensionality of temporal information remain unanswered.

In this thesis, we propose the use of "*trajectories*" to encode spatiotemporal information in fMRI data (Figure 1.1). The observation space of natural stimuli is multimodal and contains important spatial and temporal structure. For example, movie clips involve actors interacting socially and expressing various emotions. They contain dynamic sequences tied together with a unifying narrative. Such clips are likely to engage cognitive, social, and emotional networks of the brain in a time-varying manner. Thus, when individuals perceive such stimuli, they give rise to complex spatially-and-temporally varying brain responses. How do we characterize spatiotemporal information in these responses? Recurrent neural networks (RNNs)

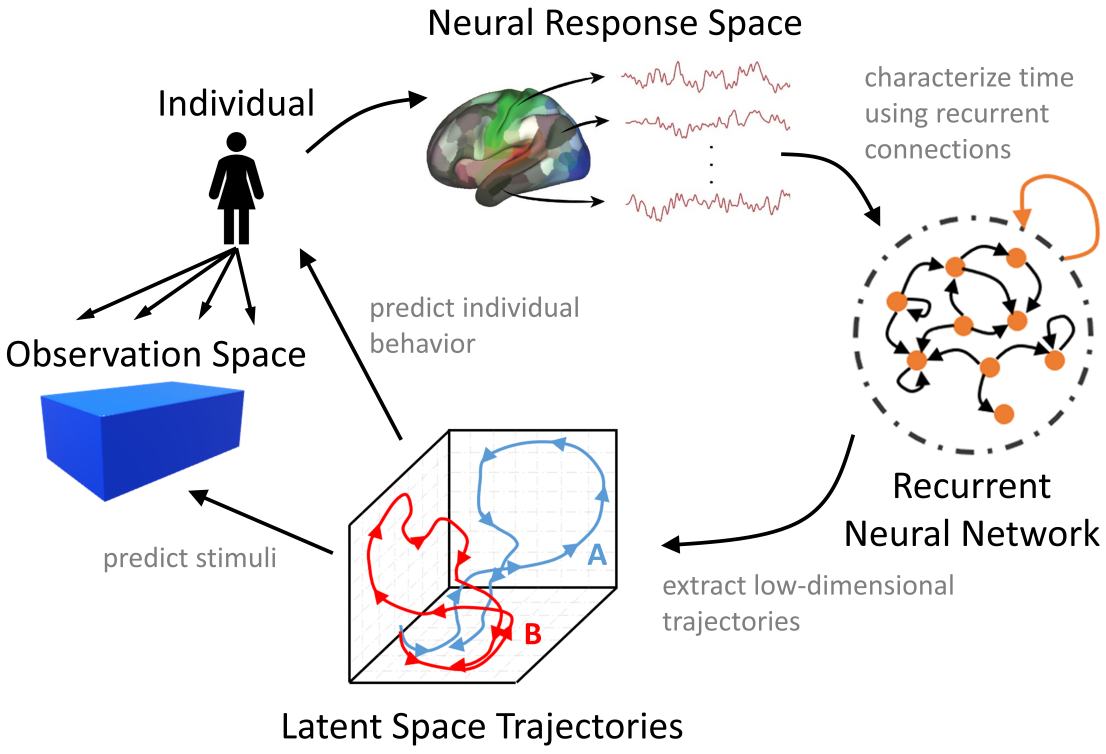

Figure 1.1: Computing with trajectories. When humans engage in diverse tasks, the observation space contains important spatial and temporal structure. Naturally, they result in neural responses that also have rich spatiotemporal dynamics. Recurrent neural networks offer a powerful framework for characterizing such dynamics. While latent representations learned by such models may be high-dimensional, important predictive information is often contained in low-dimensional representations, which we refer to as "*trajectories*". We hypothesize that such trajectories carry important information not only about the stimuli but also about the individual. For example, trajectory A and B could correspond to two different stimuli in the observation space perceived by the same individual, or to two individuals who perceive the same stimuli but have different behavioral characteristics (say, high and low working memory capacity).

have attracted considerable attention for processing temporal data [7–9]. They learn latent representations of time series data in a supervised or unsupervised manner. The latent representation at a time point is not only a function of the input at that time but also dependent on past inputs.

Although the latent space may be high-dimensional, often, low-dimensional

representations contain important predictive information. In neuroscience, research with multi-unit neuronal data has suggested that low-dimensional trajectories can be extracted from high-dimensional noisy data [3, 10]. As Yu and colleagues proposed [10], a neural trajectory potentially provides a compact representation of the high-dimensional recorded activity as it evolves over time, thereby facilitating data visualization and the study of neural dynamics under different experimental conditions (see also [11]). In this thesis, we hypothesized that low-dimensional trajectories could serve as "signatures" for task conditions or stimuli. We also hypothesized that differences in temporal evolution could be related to behavior or personality characteristics of participants.

What is the nature of explanations offered by models based on fMRI data? Firstly, the relationship between fMRI responses and neural activity has been extensively studied [12, 13]. Simultaneous recordings of fMRI and electrophysiological data in monkeys as well as fMRI and calcium imaging recordings in mice reveal that BOLD responses reflect neural activity. A key advantage of fMRI is the ability to study every region in the brain using a single cohesive model. A disadvantage is that local information at finer spatial resolution that occurs at structures within voxels are not captured. However, fMRI activity at the scale of a region or network have shown correlation with individual differences in behavior [14], and are predictive of mental illnesses [15, 16].

We also acknowledge that the low-pass nature of the blood-oxygenation response in fMRI is such that dynamics should be understood at a commensurate temporal scale (on the order of a few seconds or typically longer). Indeed, several

mental process unfold at time scales that can be captured by fMRI, such as the processing of event boundaries [17], a gradually approaching threatening stimulus [18], listening to a narrative [19], or watching a movie [20]. In summary, the spatial and temporal resolution of brain activity measured using fMRI is useful for understanding the brain [21], can provide evidence of underlying mechanisms [22], and have potential to aid clinical practice [23–25].

Part 1: Recurrent Neural Network approaches to characterizing spatiotemporal fMRI responses

In Chapter 2, we employed "*reservoir computing*", a class of recurrent neural networks, to classify fMRI data into task conditions. An input time series is fed to the reservoir, whose state changes at every time point. Reservoirs are capable of integrating current information with the past in a dynamic manner, and thus each reservoir state is a characterization of the dynamics in the data. The output layer predicts the task condition based on the reservoir state. Although reservoirs include recurrent connections similar to other RNNs, the learning component is only present in the output layer (Figure 1.2A). The reservoir weights are semi-randomly initialized. Intuitively, reservoirs are capable of separating complex stimuli because of their ability to "project" the inputs to a higher dimensional space in a context-based manner, where they are easier to classify. That is, the projection is a function of the past inputs as well as the current input.

We investigated this framework on dynamic fMRI data obtained when par-

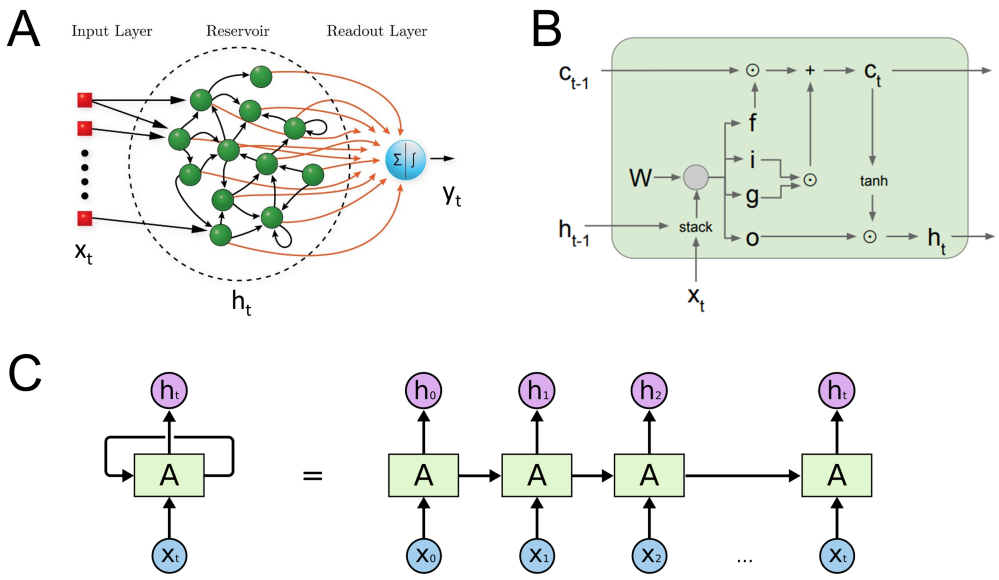

Figure 1.2: Recurrent Neural Networks (RNN). (A) In reservoir computing, recurrent weights are not learned. They are semi-randomly initialized. Only the readout layer (shown in red) is trained. (B) Depiction of a long short-term memory (LSTM) unit. The weights of this network (W) are learned during training. The gating mechanisms depicted by i, f, g, o allow for learning long-term dependencies. (C) Unrolling of an RNN [26]. Connections arriving at layers are viewed as coming from a previous time step. Fully-trained RNNs are trained using Backpropogation Through Time (BPTT). The total loss is computed by forward passing the entire time series and accumulating the loss at each time step. The gradients are then computed by passing the total loss backward through entire sequence.

ticipants observed objects interacting in either a social or random manner (in the "social" condition the objects appear to play together, for example). Initially, participants are unaware of the type of interaction. But these interactions build up after a few seconds and evolve throughout the clip.

Using reservoirs, we obtained low-dimensional representations or trajectories by a supervised selection of components that carry the most discriminative information. When examined visually, trajectories were close together at the beginning of the clips but moved further apart as the social interactions become more clear. Classification results based on the reservoir states revealed the same: accuracy was

at chance levels at the beginning of the block but increased towards the end of the clip. Although fMRI data is very high-dimensional (greater than 100,000 if we consider voxels across the whole brain), close to 12 dimensions achieved similar classification performance to the full data. Thus, trajectories captured important temporal information in the data.

Using recurrent connections, reservoirs generate a random, non-linear, high-dimensional projection of the data. To understand what aspects of the reservoir drove classification accuracy, we compared it to several approaches.

1. What if the high-dimensional projection was obtained without using random connections or non-linear units? We considered a simple concatenated model.

2. What if we build a simple autoregressive model (without the high-dimensional expansion)?

The reservoir approach outperformed the alternatives emphasizing the complexity of spatiotemporal patterns in dynamic task paradigms.

We also tested our method on data acquired when participants watched movie clips that were either "funny" or "scary", and observed similar results in terms of achieving good classification accuracy and low-dimensional representations. Finally, we identified brain regions that contributed the most to low-dimensional representations. Several regions known in the literature to be important for social cognition (in the case of classifying "social" clips) were found to contribute most to task discrimination as well. However, when data were limited to only such important brain regions, classification accuracy decreased. The results suggest that correlates of so-

cial cognition are more distributed across the brain. In conclusion, the proposed approach in Chapter 2 may provide a flexible and powerful framework to characterize dynamic fMRI information, which can be readily applied to other types of brain data, including high-density electrophysiological recordings and calcium imaging. This study has been published in *Neuroimage* [27].

"Naturalistic" stimuli such as film clips and spoken narratives resemble the complexity and dynamics of behavior and stimuli in everyday life. Neuroscience researchers are increasingly using such stimuli to complement and extend tightly controlled task paradigms [28]. The time-locked nature of these stimuli leads to a natural question: Are brain signals generated by such dynamic stimuli consistent across individuals? That is, are spatiotemporal patterns in these signals consistent and generalizable? In Chapter 3, we employed Long Short-Term Memory (LSTM) networks, another class of recurrent neural networks, to uncover distributed spatiotemporal patterns during movie-watching. Given the large temporal lengths of movie clips, the gating mechanisms in LSTMs offer a powerful framework to learn long-term dependencies (Figure 1.2B).

Unlike reservoir computing, the recurrent weights in LSTMs are learned during training using the Backpropagation Through Time (BPTT) algorithm. An LSTM can be visualized as a connected series of units, each passing data to the successor (Figure 1.2C). During training, the total loss is computed by forward passing the entire time series and accumulating the loss at each unit. The total loss is then passed backwards through the entire sequence and gradients are computed to update the weights. The availability of larger datasets enabled efficient training using

BPTT.

To understand whether watching movie clips result in spatiotemporal brain patterns that are generalizable across individuals, we trained LSTMs to predict clips. Classification accuracy exceeded 87% for 15-way classification (15 unique clips were used). But is temporal information necessary for classifying movie clips? To understand this, we systematically compared our approach to other classifiers that had varying levels of temporal modeling capability. We show that capturing short-term relationships do not suffice, and that it is crucial to capture long-term dependencies for accurate clip prediction. The results revealed that spatiotemporal patterns were distributed across time and were most effectively captured by LSTMs.

To extract trajectories associated with clips, we proposed a non-linear supervised dimensionality reduction technique. In Chapter 2, low-dimensional representations were obtained by selecting components from the reservoir latent space. However, the generation of reservoir outputs and the selection of low-dimensional components were independent of each other. Here, we employed a unified framework that simultaneously learns the best latent space for clip classification as well as the optimal low-dimensional projection. Using this technique, classification accuracy was comparable to full data revealing that important temporal properties were captured in lower dimensions.

We further investigated whether differences in temporal evolution could be linked to behavior and personality of individuals. In recent years, researchers have shifted focus from group-level inferences to characterizing single subjects [29]. A number of works have employed brain data to predict a participant's behavioral

capabilities as well as personality-based measures [30–34]. Typically, these works use a static characterization of the temporal data for prediction. Here, we predicted behavior and personality-related scores based on the learned latent representations of LSTMs. Prediction accuracy was consistently and robustly higher than existing methods on a range of behavior/personality measures.

We employed saliency maps to characterize spatiotemporally-varying brain region importance. The contribution of various regions and networks fluctuated across time but were consistent across participants. Notably, several regions with high saliency at the beginning of clips were not captured by time-averaged saliency maps. Further, brain regions that were important for predicting a particular clip were different from those important for predicting a behavioral measure using the same clip. In conclusion, our work in Chapter 3 provides further evidence that brain dynamics must be embraced for a fuller characterization of underlying processes. We believe the approach provides a powerful framework for visualizing, analyzing, and discovering dynamic spatially distributed brain representations during naturalistic conditions. This work is currently under review.

In addition, a general goal of this dissertation was to understand the benefit of techniques that can characterize temporal information in fMRI responses. Thus, we employed well-defined architecture components, such as reservoirs and LSTMs. In recent years, a number of RNN variants have emerged. For example, reservoirs have been modified to include feedback from the readout layer [35]. However, training such networks is more complicated due to stability issues [36, 37]. Among fully-trained RNNs, Gated Recurrent Units (GRU) have shown comparable performance

to LSTMs [38,39]. GRUs combine multiple LSTM gates into a single gate resulting in a simpler model. We view the framework described in Figure 1.1 as a flexible approach in which the RNN component can be exchanged as long as they prove effective in capturing spatiotemporal low-dimensional latent subspaces.

Part 2: A geometry-aware approach to comparing functional connectivity matrices

Another central goal in neuroscience is to understand the correlation structure associated with multiple brain measurements acquired across spatial locations. These correlation structures are often summarized using *functional connectivity* (FC) matrices. Such matrices have been used to understand recurring brain states, characterize dynamic functional states, perform participant identification, and for understanding how tasks modulate connectivity. In these applications, some notion of proximity or similarity between FC matrices is employed.

For example, dynamic FC analysis involves computing a time course of functional networks using limited time windows [40]. Such FCs are clustered into what are known as "states" based on k-means clustering (Figure 1.3A). The proximity measure used for clustering is the L^1 distance. The hypothesis is that the brain transitions between these well-defined states. Another application is in participant identification [41]. Given a database of FC matrices, participant identification involves labeling an unknown participant in the test database to the closest participant in the training data (Figure 1.3B). The most common technique in the literature is

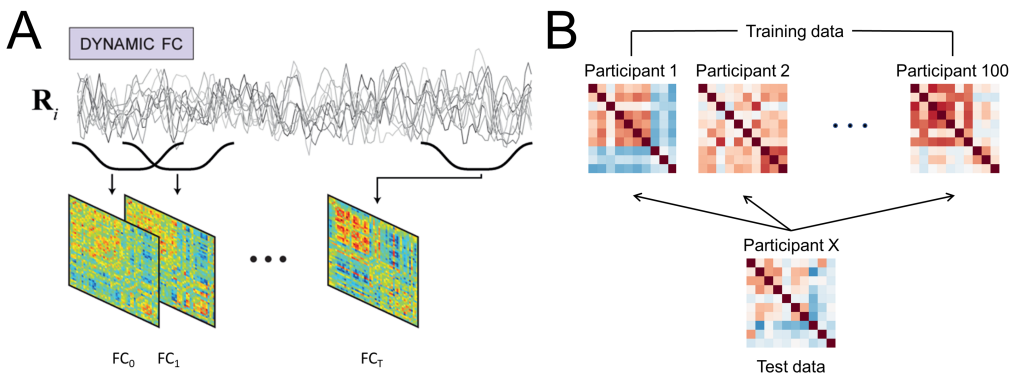

Figure 1.3: Applications of functional connectivity (FC) matrices. (A) Clustering brain networks: A time course of FC matrices is obtained using overlapping time windows. A small number of brain "states" are obtained by clustering these high-dimensional matrices using the k-means technique [40]. An L^1 distance function was employed. (B) Participant identification: A participant in the test data is labeled to the "closest" participant in the training data [41]. Pearson dissimilarity was used as the closeness measure.

to use the Pearson correlation between the upper triangular entries of the matrices themselves as the proximity measure.

However, FC matrices computed by Pearson correlating time series data are covariance matrices that lie on the *positive semidefinite cone*. Their distances must be measured along this manifold. Because of the high dimensionality of FC matrices, the choice of distance measure is particularly crucial. In Chapter 4, we show advantages of using a *geodesic distance* measure on the problem of participant identification [41]. Given fMRI data (task or resting-state), is it possible to identify a participant from her FC matrix? We observed robust improvements in both task and resting-state data. The geodesic distance measure outperformed Pearson dissimilarity even when a fourth of the data was employed to estimate the FC matrix. Further, connectivity patterns in particular subnetworks were substantially more unique than others.

We also used dimensionality reduction techniques to aid visualization of high-dimensional FC matrices. Distance information was preserved, to the extent possible, in the low-dimensional representations, and their visualizations reflected accuracy in the full-dimensional data. Such visualizations help understand the geometry of task FC structure with respect to resting-state FC. We propose that the use of the geodesic distance is an effective way to compare correlation structures of the brain across a broad range of studies including the clustering of brain states, how tasks potentially reconfigure brain networks, and to characterize inter-subject correlations. This study was presented at the *Society for Neuroscience* conference in 2019. It is also published in *Neuroimage* [42].

Chapter 2: Brain dynamics and temporal trajectories during task and naturalistic processing

Functional Magnetic Resonance Imaging (fMRI) data are acquired while participants engage in diverse perceptual, motor, cognitive, and emotional tasks. Three-dimensional images are acquired every 1-2 seconds and reflect the state of blood oxygenation in the brain, which serves as a proxy for neuronal activation. Although data are acquired temporally, they are most often treated in a quasi-static manner [43]. In blocked design experiments, fairly constant mental states are maintained for 15-30 seconds, and data are essentially averaged across multiple repetitions of a given block type, such as performing a working memory task. In event-related designs, short trials typically 1-5 seconds long are employed and the responses evoked are estimated with multiple regression.

Many fMRI studies also are constrained spatially, in the sense that activation is analyzed independently at every location in space. However, so-called multivariate pattern analysis techniques capitalize on information that is potentially distributed across space to characterize and classify brain activation [44–46]. For example, in an early study, Cox and Savoy [47] investigated the performance of a linear discriminant classifier, a polynomial classifier, and a linear support vector machine

to classify objects presented to participants from voxel activations (i.e., features) across visual cortex. Since then the field has matured and developed a wealth of approaches, including the investigation of "representational" content carried by brain signals [48]. However, given the relatively low signal-to-noise ratio of fMRI data (which necessitates a large number of repetitions of data segments of interest), the vast majority of multivariate methods for investigating brain data are "static," that is, the inputs to classification are patterns of activation that are averaged across time ("snapshots") [4]. Some studies have proposed using temporal information as well as spatial data [5, 6, 49–51]. One of the goals in such cases has been to extend the features provided for classification by considering a temporal data segment instead of, for example, the average signal during the acquisition period of interest. Despite some progress, key issues remain largely unexplored, including understanding the integration of temporal information across time, and questions about the dimensionality of temporal information (see below).

In all, despite the potential of fMRI to be used to investigate temporal structure in task data, the technique is employed in a largely static fashion. However, a fuller understanding of the mechanisms that support mental functions necessitates the characterization of dynamic properties. In this chapter, we describe an approach that aims to address this gap. At the outset, we acknowledge that the low-pass nature of the blood-oxygenation response is such that dynamics should be understood at a commensurate temporal scale (on the order of a few seconds or typically longer). Yet, many mental processes unfold at such time scales, such as the processing of event boundaries [17], a gradually approaching threatening stimulus [18], listening

to a narrative [19], or watching a movie [20].

Several machine learning techniques exist that are sensitive to temporal information. Among them, recurrent neural networks (RNNs) have attracted considerable attention [7–9]. However, effectively training RNNs is very challenging, particularly without large amounts of data ([52]; but for recent developments see [53,54]). Here, we propose to use *reservoir computing* to study temporal properties of fMRI data. This class of algorithms, which includes liquid-state machines [55], echo-state networks [56,57], and related formalisms [58,59], includes recurrence (like RNNs) but the learning component is only present in the read-out, or output, layer. Because of the feedback connections in the reservoir, the architecture has memory properties, that is, its state depends on the current input and past reservoir states. The read-out stage can be one of many simple classifiers, including linear discrimination or logistic regression, thus providing considerable flexibility to the framework. Intuitively, reservoir computing is capable of separating complex stimuli because the reservoir "projects" the input into a higher-dimensional space, making it easier to classify them. Of course, this is related to the well-known difficulty of attaining separability in low dimensions, as was recognized early on with the use of Perceptrons [60].

Reservoir computing has been effectively used for time series prediction [61], temporal signal classification [62], as well as applications in several other domains [63, 64]. Here, we show the feasibility and potential of using it for the analysis of temporal properties of brain data. The central objectives of our study were as follows. First, to investigate reservoir computing for the purposes of *classifying* fMRI data, in particular when temporal structure might be relevant, including both task

data and data acquired during movie watching. The latter illustrates the potential of the technique for the analysis of naturalistic conditions, which are an increasing focus of research. Here, classification was attempted on task condition (for example, theory of mind versus random motion) or movie category ("scary" versus "funny").

Our second goal was to characterize the *dimensionality* of the temporal information useful for classification. Many systems can be characterized by a lower-dimensional description that captures many important system properties. In neuroscience, research with multi-unit neuronal data has suggested that low-dimensional "*trajectories*" can be extracted from high-dimensional noisy data [3, 10]. As Yu and colleagues proposed [10], a neural trajectory potentially provides a compact representation of the high-dimensional recorded activity as it evolves over time, thereby facilitating data visualization and the study of neural dynamics under different experimental conditions (see also [11]). Here, we hypothesized that reservoir computing could be used to extract low-dimensional fMRI trajectories that would provide "signatures" for task conditions and/or states (Figure 2.1B). For both of our objectives, we sought to investigate them at the between-participant level (in contrast to within-participant) to ascertain the generalizability of the representations created by the proposed framework.

2.1 Methods

2.1.1 Human Connectome Project Data

We employed working memory and theory of mind data collected as part of the Human Connectome Project (HCP; [65]). Data were collected with a TR of 720 ms. We employed data from $N = 200$ participants. This included $N = 100$ unrelated participants, and a separate, non-overlapping set of $N = 100$ participants randomly selected from the $N = 1200$ data release.

2.1.1.1 Working memory dataset

Participants performed a "2-back" working memory task, where they indicated if the current stimulus matched the one presented two stimuli before, or a control condition called "0-back" (without a memory component). Data for two runs were available, each containing four 27.5-second blocks of each kind. Stimuli consisted of faces, places, tools, and body parts. To account for the cue response at the start of the block and the hemodynamic lag, data from 12-30 seconds after block onset were used (25 data points per block).

2.1.1.2 Theory of mind dataset

Participants performed a theory of mind task, where they indicated whether short video clips displayed a potential social interaction, no meaningful interaction ("random"), or they were unsure. Stimuli consisted of 20-second video clips in which

geometric objects (squares, circles, triangles) appeared to interact either meaningfully, or randomly. Data for two runs were available, each containing five video clips; thus, five social interaction and five random clips were available in total. To account for hemodynamic lag (no cue was employed), data from 3-21 seconds after block onset were used (25 data points per block).

2.1.2 Participants (movie watching)

Sixteen participants with normal or corrected-to-normal vision and no reported neurological or psychiatric disease were recruited from the University of Maryland community. Data from 12 participants (5 males and 7 females, ages 18-22 years; mean: 20.6, SD: 1.5) were employed for data analysis (two participants voluntarily quit the study before completion, and data from three participants were discarded due to head motion exceeding 4 mm). The project was approved by the University of Maryland College Park Institutional Review Board and all participants provided written informed consent before participation.

2.1.3 Movie data acquisition

Functional and structural MRI data were acquired using a 3T Siemens TRIO scanner with a 32-channel head coil. First, a high-resolution T1-weighted MPRAGE anatomical scan (0.9 mm isotropic) was collected. Subsequently, we collected six functional runs of 384 EPI volumes each using a multiband scanning sequence [66]. For 3/12 participants, the following imaging parameters were used: TR = 1.25 sec,

TE = 42.8 ms, FOV = 210 mm, voxel size: 2.0 mm isotropic, number of slices = 72, and multiband factor = 6. For the remaining 9 participants, slightly altered parameters used were: TR = 1.25 sec, TE = 39.4 ms, FOV = 210 mm, voxel size: 2.2 mm isotropic, number of slices = 66, and multiband factor = 6. For all participants, non-overlapping oblique slices were oriented approximately 20-30 clockwise relative to the AC-PC axis (2.0 mm isotropic) helping to decrease susceptibility artifacts at regions such as the orbitofrontal cortex and amygdala.

2.1.4 Movie data

We employed fMRI data collected from 12 usable participants while viewing short movie segments (duration between 1-3 minutes) with content that was either "scary," "funny," or "neutral" (neutral segments were not utilized here) (see Table A.1 for a list of the movies employed). Participants viewed one movie clip of each kind per run for a total of six runs. A total of 30 movie clips (15 of each kind) were extracted from the movie segments such that at least one clip originated from each of the movies viewed. Clips contained 25 data points (like the HCP data above), which lasted 31.25 seconds (data were acquired with a TR of 1.25 seconds). All video clips focused on parts of the movie segments that were deemed by one of the authors (M.V.) to be of high arousal/interest.

2.1.5 Preprocessing

2.1.5.1 HCP data

Data were part of the "minimally preprocessed" release, which included fieldmap based distortion correction, functional to structural alignment, and intensity normalization. Data were collected with a TR of 720 ms. We investigated cortical data which are directly provided in surface representation. In addition to the preprocessing above, we regressed out 12 motion-related variables (6 translation parameters and their derivatives) using the 3dDeconvolve routine of the AFNI package [67] (with the "ortvec" option). Low frequency signal changes were also regressed out with the same routine by using the "polort" option (with the polynomial order set automatically).

2.1.5.2 Movie data

A combination of packages and in-house scripts were used to preprocess both the functional and anatomical MRI data. The first five volumes of each functional run were discarded to account for equilibration effects. Slice-timing correction (with AFNI's 3dTshift) used Fourier interpolation to align the onset times of every slice in a volume to the first acquisition slice, and then a six-parameter rigid body transformation (with AFNI's 3dvolreg) corrected head motion within and between runs by spatially registering each volume to the first volume.

To skull strip the T1 high-resolution anatomical image (which was rotated to

match the oblique plane of the functional data with AFNI's 3dWarp), the ROBEX package [68] was used. Then, FSL's epi-reg was used to apply boundary-based co-registration in order to align the first EPI volume image with the skull-stripped T1 anatomical image [69]. Next, ANTS [70] was used to estimate a nonlinear transformation that mapped the skull-stripped anatomical image to the skull-stripped MNI152 template (interpolated to 1-mm isotropic voxels). Finally, ANTS combined the transformations from co-registration (from mapping the first functional EPI volume to the anatomical T1) and normalization (from mapping T1 to the MNI template) into a single transformation that was applied to map volume registered functional volumes to standard space (interpolated to 2-mm isotropic voxels). The overall approach described in this paragraph was based on [71] and used previously by our group [18]. The resulting spatially normalized functional data were smoothed using a 6 mm full-width half-maximum Gaussian filter. Spatial smoothing was restricted to gray-matter mask voxels (with AFNI's 3dBlurInMask). Finally, the average intensity at each voxel (per run) was scaled to 100.

2.1.6 Regions of Interest

2.1.6.1 HCP data

Because our goal was to evaluate the general framework described here, and not test specific hypotheses tied to particular brain regions, we considered cortical data only. Because for cortical data the HCP processing pipeline is oriented toward a surface representation, we employed the cortical parcellation developed by

their research group [72]. The parcellation includes 360 cortical regions of interest (ROIs), and is based on a semi-automated approach that delineates areas based on architecture, function, connectivity, and topography (see Figure A.1A).

2.1.6.2 Movie data

ROIs were determined in a volumetric fashion. To do so, we employed a simple k-means clustering algorithm that generated 500 cortical ROIs. Specifically, clustering was based on the $\{x, y, z\}$ spatial coordinates of voxels in cortex (not their time series), and an L_2 distance metric was employed to favor the grouping of nearby voxels (see Figure A.1B). We also performed our analysis with 400 and 600 ROIs and observed essentially the same results; thus, the precise choice of the number of ROIs does not appear to be critical. In addition to the cortical ROIs, given the importance of the amygdala for emotional processing in general, we also included two amygdala ROIs (one per hemisphere). Each ROI was generated by combining the lateral and the central/medial amygala (as defined in [73]) into a single region.

For both HCP and movie data, a summary ROI-level time series was obtained by averaging signals within the region.

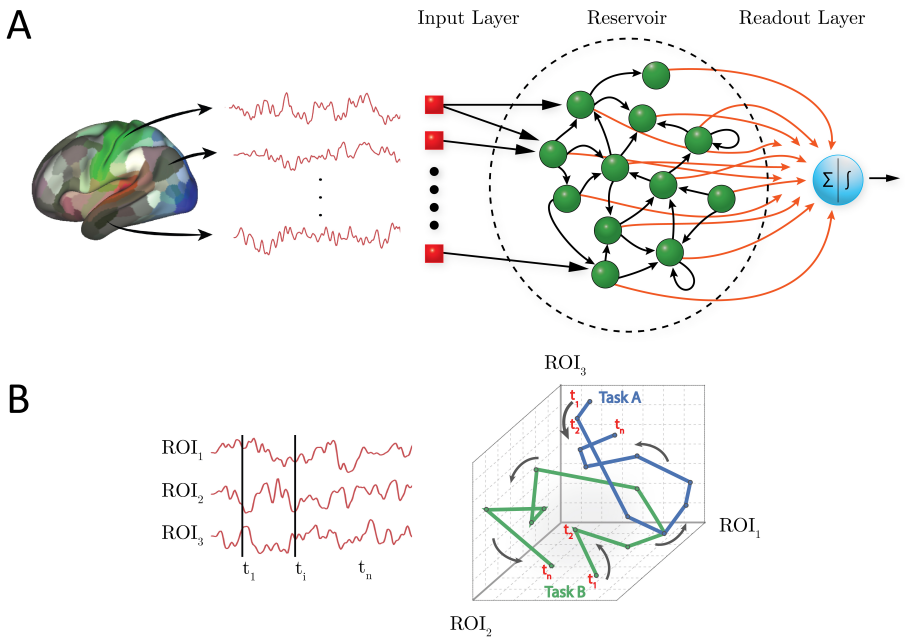

Figure 2.1: Reservoir computing and temporal trajectories. (A) Brain data are provided to a three-layer neural network. The input layer registers the input; in the present case, activation at time t across a set of regions of interest (ROIs). The reservoir layer contains units with random connections, and provides a memory mechanism such that activation at time t is influenced by past time points. The readout (output) layer indicates the category of the input; in the present case, the binary labels "0" or "1" corresponding to task/condition. Only the connections between the reservoir and the readout layer (shown in red) are adaptable. (B) Time series data can be represented as a temporal "trajectory." In the case of data from three ROIs (left), activation can be plotted along axes "1," "2," and"3" at each time point t (right). In this manner, activation during a hypothetical task exhibits a particular trajectory, whereas activation during a second task exhibits a different trajectory (blue and green lines for Task A and B, respectively). Note that the trajectories might overlap at several time points (the activation at those time points is the same for both tasks), but the entire trajectory provides a potentially unique "signature" for the task/condition in question.

2.1.7 Reservoir computing

For temporal data analysis, we adopted the reservoir formulation used in echo-state networks [56, 57]. The general reservoir computing architecture includes three main elements: an input layer, a reservoir, and a read-out (or output) layer (Fig-

ure 2.1A). The input layer registers the input and is connected with the reservoir. The reservoir contains units that are randomly interconnected within the reservoir, as well as connected to units in the read-out layer. Only connections to the read-out layer undergo learning. Here, the input layer activations, $\mathbf{u}(t)$, represented activation for the condition of interest at time t. The number of input units corresponded to the number of ROIs, and one value was input with every data sample (every time t). The output layer contained a single unit with activation corresponding to a category label (0 or 1, coding the task condition). At every time step, the activations of the reservoir units were updated, determining a reservoir state, $\mathbf{x}(t)$, and the readout, $z(t)$, was instantiated. Thus, the input time series data generated an output time series (one per time point) corresponding to category labels.

The state of the reservoir was determined by [74]

$$\tilde{\mathbf{x}}(t) = \tanh(\mathbf{W}^{\text{in}}[1; \mathbf{u}(t)] + \mathbf{W}\mathbf{x}(t-1)), \tag{2.1}$$

$$\mathbf{x}(t) = (1-\alpha)\mathbf{x}(t-1) + \alpha\tilde{\mathbf{x}}(t), \tag{2.2}$$

where $\tilde{\mathbf{x}}(t)$ is an intermediate state. The function $\tanh(x)$ was applied element-wise and implemented a sigmoidal activation function. The notation $[\cdot;\cdot]$ stands for vertical vector concatenation. Both $\tilde{\mathbf{x}}(t)$ and $\mathbf{x}(t) \in \mathbb{R}^{N_x}$, where the dimensionality of the reservoir $N_x = \tau \times N_u$ is determined by the number of input units, N_u, and the parameter τ. The dimensionality of the reservoir, N_x, is related to the memory of the reservoir, namely, the number of past data points that can influence the current output. A general rule of thumb is that for an input of size N_u, to remember τ time points in the past, the reservoir should have size at least $\tau \times N_u$

[74]. The weight matrices $\mathbf{W}^{\text{in}} \in \mathrm{R}^{N_x \times (1+N_u)}$ and $\mathbf{W} \in \mathrm{R}^{N_x \times N_x}$ are the input-to-reservoir and within-reservoir matrices, respectively. The parameter $\alpha \in (0, 1]$ is the leakage (or "forgetting") rate. Interpreting the equations above, $\tilde{\mathbf{x}}(t)$ is a function of a weighted contribution of the input plus a weighted contribution of the prior reservoir state (passed through a sigmoidal function) (Equation 1). The reservoir state, $\mathbf{x}(t)$, is a weighted average of the previous reservoir state $\mathbf{x}(t-1)$ and $\tilde{\mathbf{x}}(t)$ based on weights $(1-\alpha)$ and α, respectively (Equation 2). Overall, this reservoir formulation allows it to encode temporal information in a spatial manner, that is, across the reservoir units. The present reservoir implementation utilized code from the Modeling Intelligent Dynamical Systems research group (`http://minds.jacobs-university.de/research/esnresearch/`).

A key idea in reservoir computing is that the weight matrices $\mathbf{W}^{\text{in}}$ and $\mathbf{W}$ are not trained, but instead generated randomly (unlike RNNs which include adaptable weights in all layers). The non-symmetric matrix $\mathbf{W}$ is typically sparse with nonzero elements obtained from a standard normal distribution, $\mathcal{N}(0, 1)$; here, of the $N_x \times N_x$ matrix entries, $10N_x$ were randomly chosen to be non-zero. The input matrix $\mathbf{W}^{\text{in}}$ is generated according to the same distribution, but typically is dense. It is crucial to ensure that the largest absolute value of the eigenvalues of the reservoir weight matrix $\mathbf{W}$ be less than 1, as this ensures the "echo state" property [56]: the state of the reservoir, $\mathbf{x}(t)$, should be uniquely defined by the fading history of the input, $\mathbf{u}(t)$.

2.1.8 Classification

The reservoir state, $\mathbf{x}(t)$, can be viewed as a random non-linear high-dimensional expansion of the input signal, $\mathbf{u}(t)$. If the inputs are not linearly separable in the original space $\mathbb{R}^{N_u}$, they often become separable in the higher dimensional space, $\mathbb{R}^{N_x}$, of the reservoir. Such so-called "kernel tricks" are common in machine learning algorithms [75, 76], and reservoirs embed that property within a temporal processing context.

The read-out layer of a reservoir architecture can employ one of multiple simple components, including linear or logistic regression, or support vector machines. Here, we employed ℓ_2-regularized logistic regression with a constant inverse regularization parameter, $C = 1$ [77], for two-class classification. Given a set of data points and category labels, a logistic regression classifier learns the weights of the output layer, $\mathbf{W}^{\text{out}}$, by maximizing the conditional likelihood of the labels given the data. A gradient descent algorithm searches for optimal weights such that the probability $P[z(t) = 1|\mathbf{x}(t)] = \sigma(\mathbf{W}^{\text{out}}\mathbf{x}(t))$ is large when $\mathbf{x}(t)$ belongs to class "1" and small otherwise; $\sigma(s) = \frac{1}{1+\exp(-s)}$ is a logistic function. The classes considered here were "2-back" vs. "0-back" for working memory, "social" vs. "random" for theory of mind, and "scary" vs. "funny" for movie clips.

Because we were interested in temporal properties, classification was performed at every time t, with a single classifier. Thus, as stated above, the input time series data generated an output time series corresponding to category labels, $\mathbf{z}(t)$.

Finally, note that our objective was to characterize the capabilities of the

reservoir framework to capture temporal information for classification as a function of time. Accordingly, we employed the "minimal" classification machinery at the output end of our algorithm. Had the objective been to maximize classification values, we could have included, for example, a "second classifier" (that is, one after the readout layer) that considered simultaneously all classification values $\mathbf{z}(t)$ during the block, for example.

2.1.9 Dimensionality reduction

Functional MRI data are very high-dimensional if one considers all the voxels or surface coordinates acquired with standard imaging parameters. Typical anatomical parcellations considerably reduce the dimensionality as 100 to 1,000 ROIs are usually employed (and one time series is commonly employed per ROI). Whereas this represents a major reduction in dimensionality, it is important to understand if lower-dimensional characterizations of the data are informative. Here, we sought to determine classification accuracy of temporal fMRI data of lower-dimensional representations. In particular, what is the lowest dimensionality that provides performance comparable to that obtained with the "full" dimensionality? Recall that because reservoir states, $\mathbf{x}(t)$, are non-linear high-dimensional expansions of the input signals, $\mathbf{u}(t)$, their dimensionality is higher than the number of ROIs (by the factor τ; see above).

For dimensionality reduction, we employed principal component analysis (PCA) to the reservoir states, $\mathbf{x}(t)$ (Figure 2.2A). In brief, PCA provides a coordi-

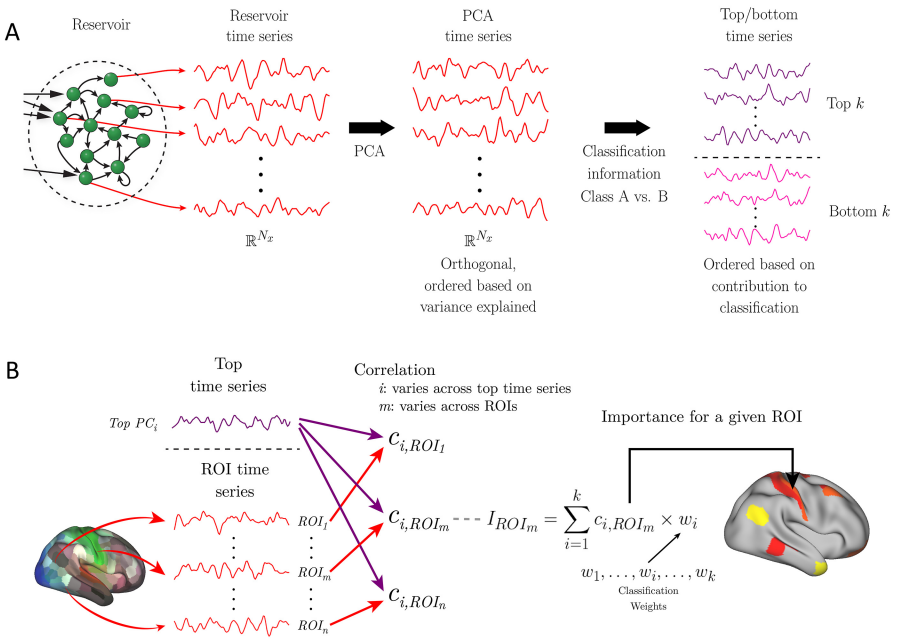

Figure 2.2: Dimensionality reduction and brain activation. (A) Reservoir activation provides a high-dimensional "expansion" of the input vectors at every time t. In this manner, the reservoir is associated with a reservoir time series of dimensionality $N_x = \tau \times N_u$, where τ is a parameter and N_u is the size of the input vector (here, the number of ROIs employed). The first step of dimensionality reduction employed principal component analysis (PCA). Subsequently, the dimensions were ordered based on the weights of the logistic regression classifier (the larger the absolute value of the weight, the more important the dimension for classification). We refer to the data along these dimensions as "top" (indicative of one output category) and "bottom" (indicative of the alternative category) time series. (B) To indicate brain regions expressing top time series information, each top time series (at left, only one is shown for simplicity) was correlated with the original fMRI time series of each ROI. The correlations along with the weights associated with the top time series are indicative of the *importance* of an ROI to classifying a condition (as the active condition). A set of ROIs can then be indicated on the brain (right) that express the k top time series based on the importance values, I_{ROI}. For example, the l ROIs with largest importance values can be shown, or those ROIs such that the importance exceeds a specific threshold. Taken together, although the reservoir time series representation is a high-dimensional expansion of the input data, it is possible to map the brain regions that most express the top time series, which are the ones providing the greatest contribution to classification.

nate transformation such that the dimensions are orthogonal. In the new coordinate system, the transformed reservoir state, $\mathbf{y}(t)$, has the same dimensionality as the

original representation. It is possible to reduce the dimensionality of the input by retaining a subset of the dimensions that capture the most variance of the original signals. Our goal, however, was to perform dimensionality reduction while considering dimensions that were useful for classification, and not necessarily capturing the most variance. To do so, we performed logistic regression analysis using PCA-transformed states, $\mathbf{y}(t)$, and used the weights of the resulting classifier to select the principal components that best distinguished the task conditions (somewhat akin to partial least squares; see [78]). Components associated with large positive weights encourage the decision toward one of the classes, whereas those associated with large negative weights encourage the decision toward the other class. We can then select the k dimensions with largest positive weights and the k dimensions with the largest negative weights, which we called the "top" and "bottom" principal components; we called time series data along the k dimensions "top and bottom time series." For example, the dimension with the largest positive weight (call it dimension 1) is associated with time series $y_1(t)$. See (Figure 2.2A) for a schematic of the sequence of data transformations. Importantly, since these components were determined by using classifier weights, which were based solely on training data, test data were unseen and could be used to assess classification performance (see Section 2.2.2).

2.1.9.1 Region importance

The high-dimensional representation of the reservoir, or the lower-dimensional representation of the k top/bottom components, is considerably re-

moved from the original fMRI time series. However, it is important to determine which original ROI time series express the most information about them, what we call *region importance*. To do so, we first computed the Pearson correlation between each original ROI fMRI time series and each of a number of top time series. To facilitate interpretation of importance, we employed only top time series because they contributed positively to classification performance, that is, they had positive classification weights (Figure 2.2A); recall that positive weights provided evidence for the "active class" and negative weights for the control condition.

The contribution of an ROI to classification was not only dependent on its correlation with a top time series but also the logistic regression weight associated with the time series. Specifically, the weight w_i from the PCA-transformed reservoir dimension, $y_i(t)$. Thus, the "importance value" of an ROI to a particular task condition was based on the correlation value times the classification weight (Figure 2.2B). Finally, an importance index for an ROI was obtained by adding the extent to which an ROI time series "loaded" (correlated with) onto k top time series corresponding to the task (k was 5 for working memory data, 6 for theory of mind data, and 6 for movie data; see Results for explanation of how k was determined). Importance values were then shown on brain maps (for illustration, we display the 25 highest importance values/ROIs on the brain). For display of importance across tasks, values were rescaled into the range $[0, 1]$: $I'_{\text{ROI}} = \frac{I_{\text{ROI}} - \min}{\max - \min}$, where I_{ROI} is the importance value prior to rescaling.

2.1.10 Additional temporal analyses

To understand the ability of reservoirs to integrate information across time, temporal information was also used in a straightforward manner. Here, the activations across a block were concatenated into a single long vector of size `number-of-ROIs` $\times$ `number-of-time-points`. The resulting vector was then used as input to a logistic regression classifier (instead of data at each time step separately) and performance determined.

To assess the role of the non-linear expansion in the reservoir, we compared the results with those obtained with a linear autoregressive model, a standard technique used to model time series data. Activations at time t for each ROI k, $u_k(t)$, were predicted based on the previous p time points, such that the predicted value at time t was given by

$$\hat{u}_k(t) = \beta_0 + \sum_{l=1}^{p} \beta_l u_k(t-l), \tag{2.3}$$

where p is the so-called model order. The estimated coefficients, β_i, that minimize the squared error between $u_k(t)$ and $\hat{u}_k(t)$ can be obtained via least squares. As routinely done, the first p time points in the block were ignored in this AR(p) model. The activations predicted based on this model were used to train a logistic classifier, as done with reservoirs.

2.1.11 Statistical approach and tests

2.1.11.1 Studying reservoir parameters

Our initial goal was to investigate the ability of reservoirs to capture temporal information in fMRI data. Accordingly, we varied the parameters α (forgetting rate) and τ (ratio of the number of reservoir to input units), which together determine the memory properties of the reservoir. To determine classification accuracy, we employed a *between-subject* cross-validation approach. For HCP data, $N = 100$ unrelated participants were used (for reference, we will call this the "first" dataset). Five-fold cross-validation was employed by randomly splitting the data into 80-20 train-validation sets: in each fold, 80 participants were used to train the reservoir, and 20 participants for validation (that is, to determine classification accuracy in unseen data). This procedure was applied for each of the $\alpha \times \tau$ parameter combinations.

Because we were interested in temporal properties, classification was performed at every time t. Classification accuracy for a block was defined on the time point with the best classification accuracy, t_{best}, during the block. We did not employ the average accuracy across the entire block, because for temporally varying data some segments of the block would not be expected to contain distinguishing information; for instance, the beginning of a block (see Figure 2.6). To improve robustness, we considered t_{best} and its two adjacent time points, $t_{\text{best}} - 1$ and $t_{\text{best}} + 1$, such that accuracy was the average across these three time points. Note that t_{best} was

defined on training data only and applied on test data that was not used to define it. Overall, the "first" dataset served to investigate reservoir parameters and define the best-performing α, τ, and t_{best}.

To evaluate the classification accuracy of reservoirs, we employed permutation testing [79]. Given the computational demands of permutation testing in our framework, p-values were based on 1000 iterations (with the exception of the test of randomizing temporal information; see below). The best-performing reservoir parameters were used to train a logistic classifier (see Section‘2.1.8) by utilizing the entire $N = 100$ participants of the "first" dataset, but accuracy was determined entirely based on a separate $N = 100$ dataset (for reference, the "second" dataset). This ensured that classification information generalized to completely unseen data. The observed accuracy was then compared to a null distribution of accuracy that was obtained by repeating this procedure 1000 times but with class labels randomly permuted; for each iteration, training with permuted labels was performed on the "first" dataset and testing was based on the "second" dataset. If m is the number of iterations where the classification accuracy on data with permuted labels exceeded the accuracy on data with true labels, and k is the total number of iterations, the p-value was obtained as $p = \frac{m+1}{k+1}$.

2.1.11.2 Comparison with other methods

We compared the performance observed with reservoirs to three other methods. The first was to simply test classification on raw activation signals. In this case, the

logistic classifier was directly fed the inputs $\mathbf{u}(t)$; everything else was identical to the classification with reservoirs. In other words, the inputs to classification were directly from the input layer and not the reservoir (see Figure 2.1A). Thus, identical to the case of reservoirs, classification on activation signals generated a time series of corresponding labels $z(t)$. The other two methods employed temporal information as outlined previously: concatenating data across time points in a block, and using autoregressive modeling. The reservoir used the best-performing α, τ, and t_{best} obtained using the "first" dataset. Likewise, the order of the autoregressive model ($p = 10$) was the best performing one obtained with the "first" dataset; the model orders investigated were $p = \{2, 5, 10\}$, which were comparable to the reservoir parameter τ values, but for results as a function of p, see Figure A.2). The actual comparison between methods was established based on the "second" dataset. To compare accuracy values, a Wilcoxon signed-rank test was utilized.

2.1.11.3 Randomizing temporal information

To test whether the temporal order within a block is informative, data points within a block were randomly shuffled. For fMRI data, simply reshuffling breaks the serial dependency in the data, and so a "wavestrapping" approach was used [80]. In this manner, the autocorrelation structure is preserved by shuffling the wavelet coefficients at each level (which are whitened and therefore exchangeable). Given the computational demands of the procedure, the associated permutation testing was based on 100 iterations.

2.1.11.4 Movie data

For movie data, we only had a limited amount of data. Accordingly, all classification accuracy results were based on 6-fold cross-validation by randomly splitting the data into 10-2 train-validation sets (10 participants for training, 2 participants for testing).

2.2 Results

Initially, we employed Human Connectome Project (HCP) data from two tasks: working memory and theory of mind. Working memory was chosen to represent a task with a relatively stable "cognitive set" (at the time scale of fMRI). For this case, the active condition comprised 25-second blocks of the so-called 2-back memory task, where participants were asked to indicate if the current item matched the one before the immediately preceding one. We employed the 0-back condition as a comparison condition (no working memory requirement). In contrast to working memory, theory of mind data were expected to exhibit some form of dynamics. During the active condition, participants watched 20-second clips containing simple geometrical objects (including squares, rectangles, triangles, and circles) that engaged in a socially relevant interaction (for example, they appeared to initially fight and then make up) that unfolded throughout the duration of the clip. When watching such clips, one has the impression that the potential meaning of the interactions gradually becomes clearer. The baseline condition in this case consisted of same-duration clips of the same geometrical objects following random motion.

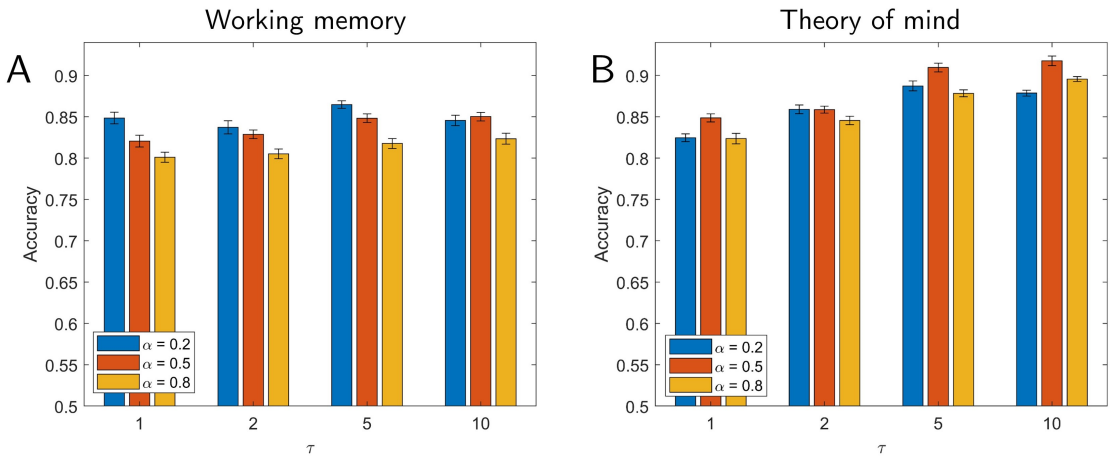

Figure 2.3: Classification accuracy for working memory (A) and theory of mind (B) tasks. Parameters varied included α (which determines the forgetting rate) and τ (which determines reservoir size), both of which influence the memory properties of the reservoir. Performance did not vary substantially as a function of reservoir memory for the working memory task (A) but improved as memory increased for the theory of mind task (B). Error bars show the standard error of the mean across validation folds.

To investigate the ability of the reservoir to capture temporal information in fMRI data, we varied the parameters α (forgetting rate) and τ (ratio of the number of reservoir to input units), which together determine the memory properties of the reservoir. Accuracy in classifying theory of mind task increased as the size of the reservoir increased, and exceeded 90% (Figure 2.3B), which robustly differed from chance (permutation test, $p < 10^{-3}$). In contrast, for the working memory task, accuracy differed from chance (permutation test, $p < 10^{-3}$) but remained essentially the same, showing that enhanced performance was not always simply due to an increase in reservoir size (Figure 2.3A).

We reasoned that if temporal information and context are important, classification should be affected by temporal order, especially in the case of theory of

mind data. To evaluate this claim, we trained the classifier without temporal information, namely, by randomly shuffling the data points in a block prior to training (while preserving autocorrelation structure; see Methods); testing was performed with blocks that were temporally ordered. In this case, mean classification accuracy was drastically reduced to 67.1% correct (using the same best-performing parameters) compared to 91.9% correct (permutation test, $p < 0.01$). This further indicates that it was not simply the high-dimensional expansion of the reservoir but also its memory that helped improve classification. For completeness, we also tested classification of working memory data in the same manner. In this case, mean classification accuracy was 74.4%, which was a relatively small (but robust; permutation test, $p < 0.01$) decline in performance relative to the best mean accuracy of 86.3% on the unshuffled working memory data.

2.2.1 Comparisons with other approaches

To better characterize the classification performance of reservoirs, we performed a series of comparisons with simpler schemes. All results in the present section were obtained by evaluating the "second" dataset and are summarized in Table 2.1. Classification accuracy using raw activation data (no reservoir, that is, $\mathbf{u}(t)$ signals) was 77.6% for working memory and 84.2% for theory of mind. For theory of mind, when the reservoir size was small ($\tau = \{1, 2\}$), accuracy was comparable to that with raw activation (see Figure 2.3B). It appears that when the number of reservoir units is relatively small, the reservoir representation of the data

Working memory: "2-back" vs. "0-back"					
	Accuracy		Accuracy	p-value	Signed-rank
Reservoirs	86.3%	Raw activation	77.6%	$< 10^{-8}$	821
		Concatenation	82.8%	$< 10^{-3}$	761
		Autoregressive model	81.1%	$< 10^{-3}$	1432
Theory of mind: "social" vs. "random"					
	Accuracy		Accuracy	p-value	Signed-rank
Reservoirs	91.9%	Raw activation	84.2%	$< 10^{-8}$	2062
		Concatenation	86.6%	$< 10^{-3}$	1370.5
		Autoregressive model	87.8%	$< 10^{-3}$	1172
Movie data: "scary" vs. "funny"					
	Accuracy		Accuracy	p-value	Signed-rank
Reservoirs	70.9%	Raw activation	60.2%	$< 10^{-3}$	77
		Concatenation	65.3%	0.0151	69.5
		Autoregressive model	64.6%	0.0107	70

Table 2.1: Classification accuracy for reservoirs and additional processing approaches. The p-values were determined via Wilcoxon signed-rank tests comparing classification accuracy of reservoirs to each method.

is poor, particularly when the forgetting rate is high, possibly due to the inability of the reservoir to generate a satisfactory representation of dynamically changing data with fewer dimensions.

The next two approaches explicitly considered temporal properties of the data. First, we concatenated activation signals from multiple time steps, and performed logistic classification on the concatenated data. Classification accuracy of working memory data was 82.8% correct and of theory of mind data was 86.6% correct. Next, we sought classification with an autoregressive model, which yielded accuracy of 81.1% correct (working memory) and 87.8% correct (theory of mind) (both of which were obtained with a model of order $p = 10$). Although performance with these

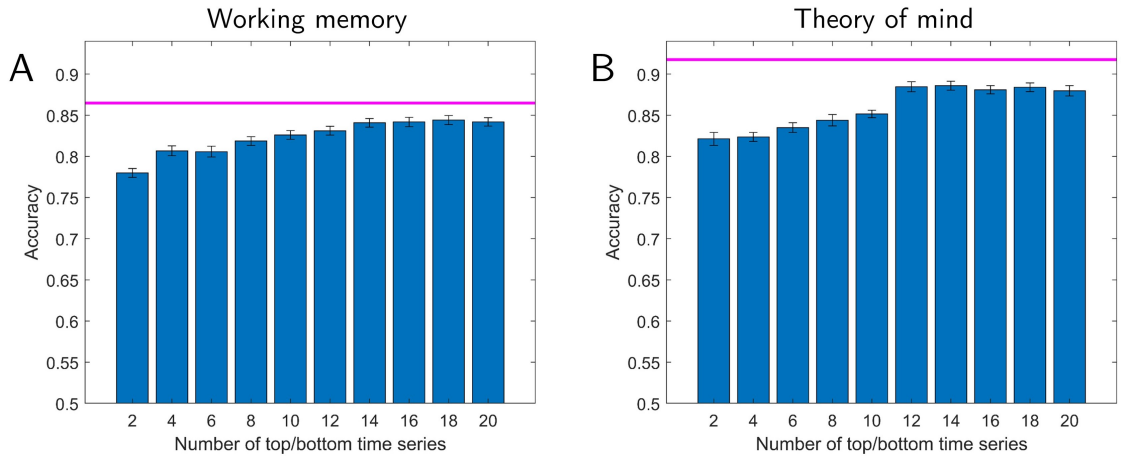

Figure 2.4: Lower-dimensional representation of reservoir signals and classification accuracy. Accuracy is shown as a function of the number of top plus bottom components. The magenta line indicates the highest performance using all components. Classification accuracy with a lower-dimensional representation reached within 95% of the the full data with 10 and 12 dimensions for working memory (A) and theory of mind (B), respectively. Error bars show the standard error of the mean across validation folds.

two methods was relatively close to that with reservoirs, the latter was consistently superior (see Table 2.1). Finally, note that the classification values across methods used in the present study were rather stable, as illustrated by the comparison of the estimates based on cross-validation ("first" dataset) and those of the "second" dataset (Table A.2); recall that all statistical results were based entirely on the "second" dataset which was never used for parameter selection.

2.2.2 Low-dimensional representation

We sought to investigate the dimensionality of the reservoir representation capable of classifying fMRI data. To do so, we performed PCA on reservoir data and determined the number of principal components required to achieve classification

performance similar to that on the full data. Instead of considering components in terms of the variance explained, we considered "top" and "bottom" components based on how they improved classification (see Methods; Figure 2.2A). Figure 2.4 shows classification accuracy as the number of components was increased from 2 to 20 in steps of 2 (one top and one bottom component were added together at a time). For working memory, 10 principal components (5 top and 5 bottom) were required to attain classification at 95% of the level of the full dimensionality; for theory of mind, 12 principal components (6 top and 6 bottom) were required. Note that these components captured only 7% and 8% of the total variance of the working memory and theory of mind datasets, respectively, which should be compared to 72% and 71% captured by first 10 and 12 components when they were selected based on the amount of variance explained (and not classification), consistent with the idea that a relatively small percentage of the original signal variance was informative for classification.

Figure 2.4 also shows that classification with only the top/bottom 2-4 components attained accuracy at approximately 90% of that obtained with the full dimensionality. We could thus capitalize on this property and select three components so as to visualize their trajectories as a function of time (Figure 2.5). For working memory data, the trajectories indicated that the two conditions should exhibit better-than-chance classification even at the beginning of the block. In contrast, for theory of mind data, the trajectories of the social and random conditions initially overlapped, but later became quite distinct. To qualify these observations, we plotted classification accuracy as a function of time during task blocks (for various

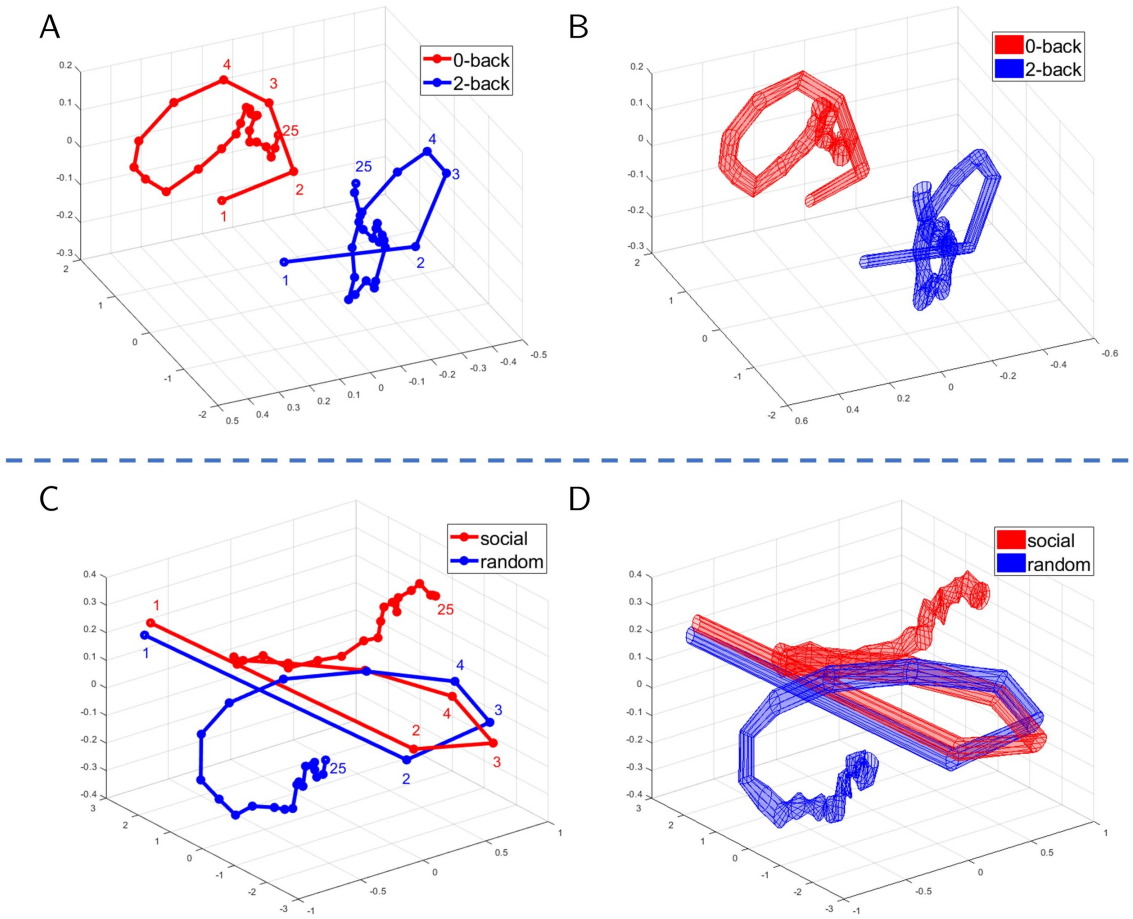

Figure 2.5: Temporal trajectories for task fMRI data. Mean trajectories are displayed in (A) for working memory and (C) for theory of mind. Variability (standard error across participants) is displayed in (B) and (D), respectively. For working memory data (A-B), the trajectories were well separated throughout the block. For theory of mind data (C-D), the trajectories initially overlapped but diverged after 6-7 points. Trajectories were based on the top three principal components.

reservoir configurations). Figure 2.6 shows the results for the full dimensionality; results for the top/bottom components are displayed in Figure A.3. For working memory, accuracy was initially around 70% correct, and increased gradually up to 85% for the best reservoir configuration. For theory of mind, accuracy was initially at chance, and increased more abruptly between time points 5-8 (3.5-5.5 seconds),

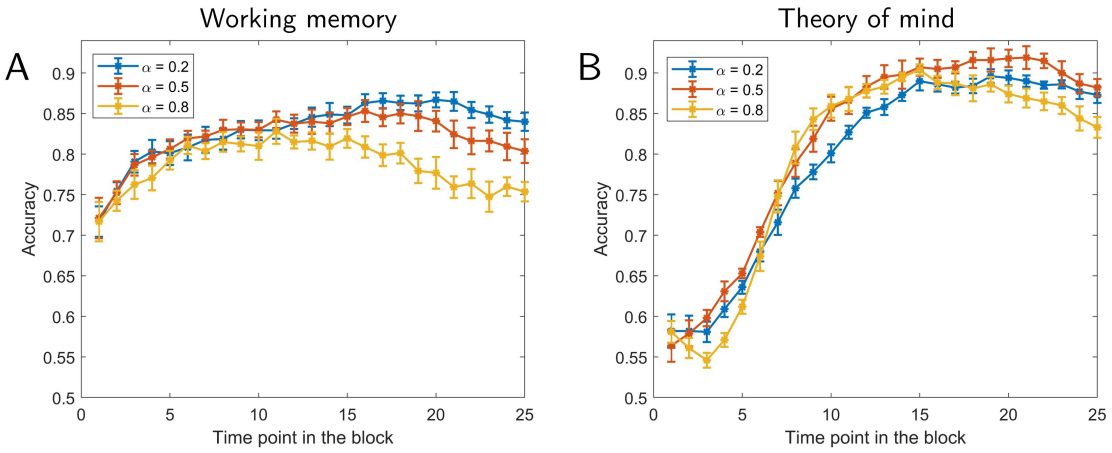

Figure 2.6: Classification accuracy as a function of time. Results for working memory (A) and theory of mind (B). Accuracy is shown as a function of time point within a task block. Different curves show results for different forgetting rates , α. The values of τ were based on the parameters exhibiting highest accuracy in Figure 2.3. Error bars show the standard error of the mean across validation folds.

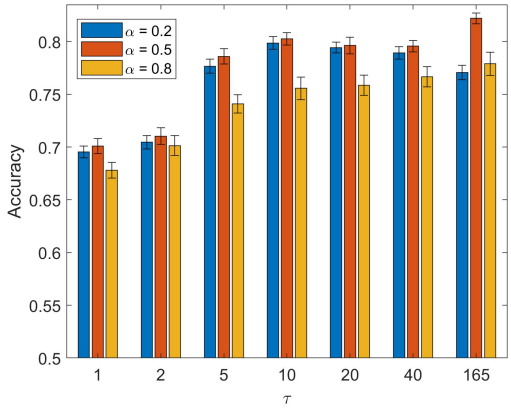

Figure 2.7: Classification accuracy for theory of mind data with regions involved in "social animation" (based on a meta-analysis of fMRI studies). Performance largely leveled off with a parameter $\tau = 5$ or larger . Because the meta-analysis only included 22 regions of interest, we increased τ so as to match the reservoir size with that used with the full dimensionality ($\tau = 165$). Note that accuracy was limited to around 80% even when the reservoir was large, indicating that the limiting factor was not the size of the reservoir. Error bars show the standard error of the mean across validation folds.

eventually attaining classification over 90%.

How does classification performance based on a lower-dimensionality represen-

tation compare to that obtained with regions previously reported to be engaged by theory of mind? To investigate this issue, we used ROIs from a meta-analysis of prior fMRI studies [81], and selected those found to be engaged during social animation tasks. The results based on the 22 ROIs from the meta-analysis are shown in Figure 2.7. Performance mostly leveled off with $\tau = 5$ at around 80% correct. Note that this performance was lower than that observed with the top/bottom 10 dimensions by about 5%. It is also instructive to compare the results obtained with the meta-analysis ROIs to those with the full data, with the latter exhibiting classification accuracy about 10% higher. The results with the the meta-analysis ROIs did not change appreciably even when the size of the reservoir was increased to match the much larger reservoir size used with the full dimensionality (this was the case when $\tau = 165$; recall that the size of the reservoir is given by τ times the size of the input vector). Thus, inferior performance with meta-analysis ROIs was not simply due to the size of the reservoir.

Finally, we also investigated the low-dimensional representation obtained using principal components directly based on activation signals (Figure 2.8). For working memory, a small number of components (3 top and 3 bottom) attained classification at 95% of the level of full activation data. However, for theory of mind, 28 components (14 top and 14 bottom) were required. This was more than twice of what was required for the reservoir data indicating that they captured more information required for classification in fewer dimensions. For completeness, Figure A.4 shows temporal trajectories when principal components were based on activation data; it appears that these do not provide temporal signatures as informative as

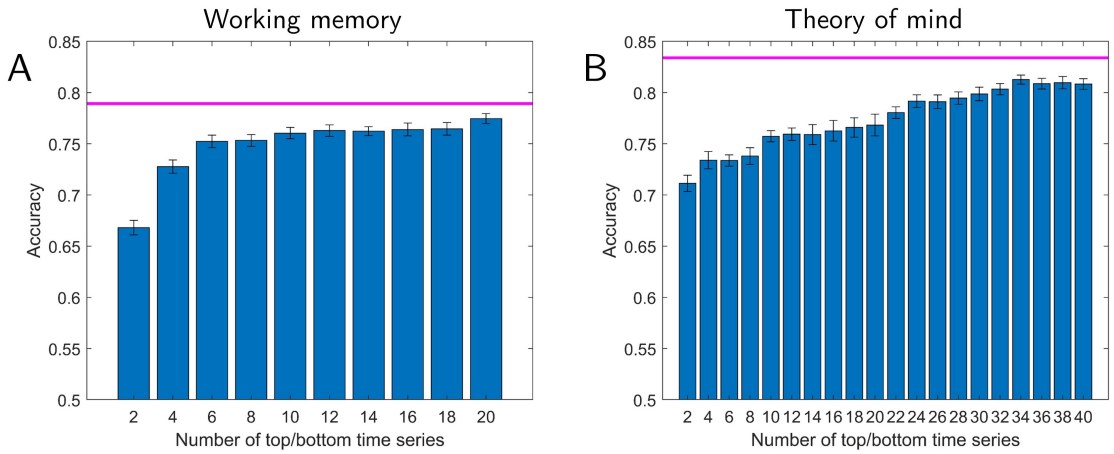

Figure 2.8: Lower-dimensional representation of activation data (no reservoir) and classification accuracy. Accuracy is shown as a function of the number of top plus bottom components. The magenta line indicates the highest performance using the full-dimensional activation data. For working memory (A), classification accuracy with the lower-dimensional representation reaches 95% of the full data with 6 dimensions. However, for the theory of mind (B), 28 dimensions (as opposed to 12 when using reservoir data) of the lower-dimensional representation. Error bars show the standard error of the mean across validation folds.

with reservoirs.

2.2.3 Mapping low-dimensional representations to the brain

We sought to determine the brain regions providing the greatest contributions to classification (see also [82]). To do so, we computed an importance index for each ROI based on time series data (see Figure 2.2B). Figure 2.9 illustrates some of the ROIs supporting classification for the working memory and theory of mind tasks selected based on the highest importance values. For this analysis, we used the top 5 time series for working memory and top 6 for theory of mind (the top components that were part of the those attaining 95% classification accuracy relative to the full dimensionality, as discussed in the previous section).

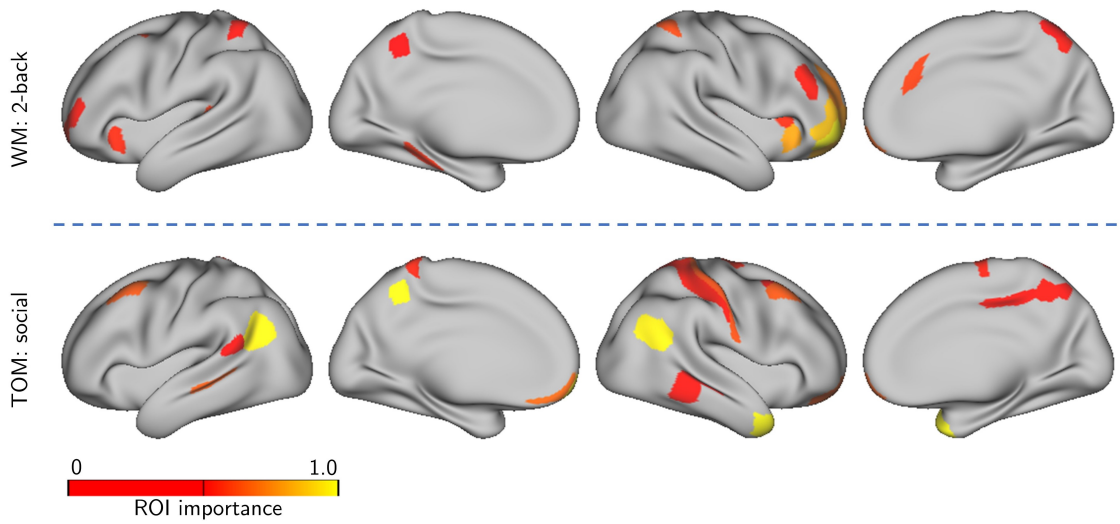

Figure 2.9: Importance maps for task data. Lower-dimensional time series representation expressed on the brain. The colored ROIs are those with original fMRI time series expressing ("loading on") "top" time series the most (see Figure 2.2 for details).(A) Regions supporting "2-back" in the classification of working memory data. (B) Regions supporting "social" in the classification of theory of mind data.

2.2.4 Movie clips

We further investigated our framework by attempting to classify data segments extracted from movies (31.25 seconds long). Twelve usable participants viewed short movie clips (between 1-3 minutes long; see Methods) of scary or funny content. Given the emotional content of the clips, we added left and right amygdala ROIs to the set of cortical ones. Classification accuracy ("scary" vs. "funny" clips) is displayed in Figure 2.10A and reached around 70% correct for larger reservoirs (which was robustly above chance levels; permutation test, $p < 10^{-3}$). Like in the case of theory of mind data, performance improved with larger reservoirs. The accuracy for individual movies was between 60% and 80%, showing that classifier performance

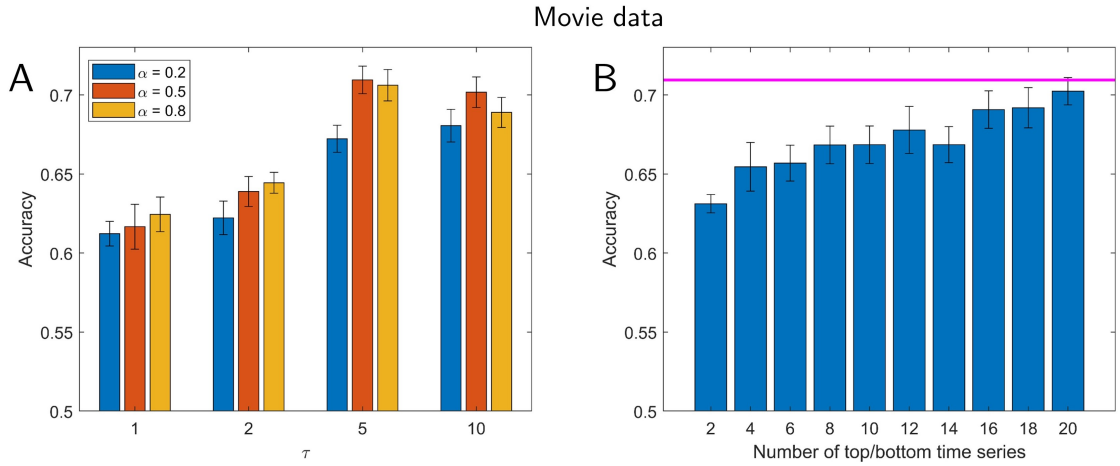

Figure 2.10: Classification for movie clips. Participants viewed short movie clips that were either scary or funny. (A) Accuracy as a function of the parameters α (which determines the forgetting rate) and τ (which determines reservoir size). (B) Lower-dimensional representation and classification accuracy. Accuracy is shown as a function of the number of "top" plus "bottom" components. The magenta line indicates the highest performance using all components. Classification accuracy with a lower-dimensional representation reached that of the full data with around 20 dimensions, and reached within 95% of the the full data with 12 dimensions. Error bars show the standard error of the mean across validation folds.

was not driven by one or a few of the movies watched. In addition, we compared classification with reservoirs with that obtained with activation signals (no reservoir; 60.2%), concatenated data (65.3%), and an autoregressive model (64.6%). As in the case of task data, reservoirs performed best, although the numerical difference was relatively modest (Table 2.1).

We also investigated lower-dimensional representations of movie data (Figure 2.10B). Classification accuracy with 20 dimensions (out of 502) performed at the same level as with the full dimensionality, and with 12 dimensions within 95% of that with all dimensions. In a more exploratory fashion, we investigated temporal trajectories during movie watching. We compared trajectories generated from indi-

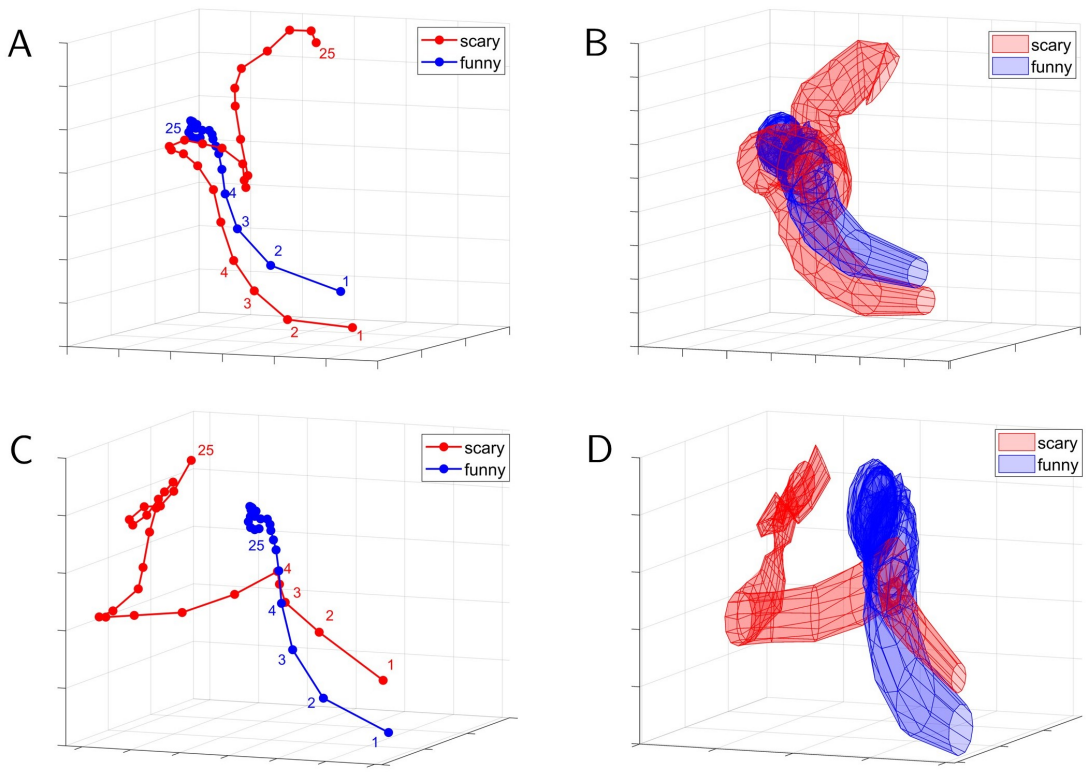

Figure 2.11: Temporal trajectories for sample clips in the movie data. The red trajectories are for particular "scary" movie clips whereas the blue trajectories are averaged across all "funny clips". Mean trajectories are displayed in (A) and (C) for two particular scary movie clips. Variability (standard error across participants) is displayed in (B) and (D), respectively. For the movie clip in (A), the trajectories started to separate later than they do for the movie clip in (C). An analysis of the accuracy of these clips as a function of time revealed similar properties. Trajectories were based on the "top" three principal components.

vidual scary clips with the average trajectory observed for funny movie segments. Some scary clips exhibited trajectories that diverged from the mean trajectory for funny clips earlier on, whereas some diverged later in time (Figure 2.11), properties that were also apparent in the time course of classification accuracy values. To determine brain regions that most contributed to classification, we computed the importance index for each ROI as with task data. Figure 2.12 illustrates some of

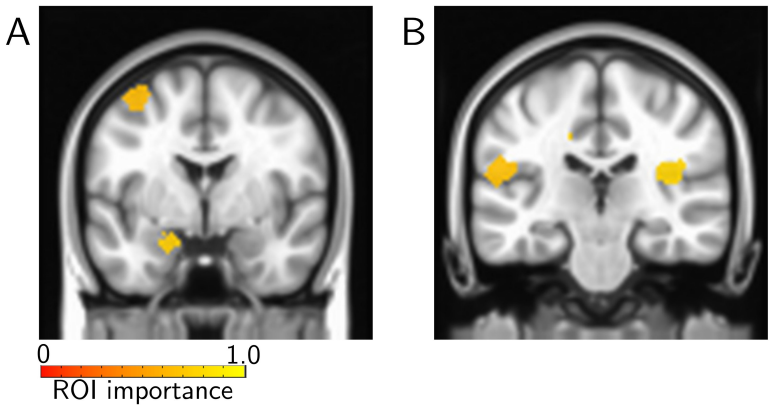

Figure 2.12: Importance maps for movie data. Lower-dimensional time series representation expressed on the brain. The colored ROIs are those with original fMRI time series expressing ("loading on") "top" time series the most (see Figure 2.2 for details). Regions supporting "scary" included the left amygdala (A) and the insula (B).

the brain regions involved when we used the top 6 time series, which attained 95% classification accuracy relative to the full dimensionality (see Figure 2.10).

2.3 Discussion

In this chapter, we sought to analyze fMRI data with reservoir computing which, like recurrent neural networks, is a technique developed to process temporal data. We show that reservoirs can be used effectively for temporal fMRI data, both for classification and for characterizing lower-dimensional trajectories of temporal data. Importantly, the approach was performed in an out-of-sample fashion, namely, performance was only evaluated in participants whose data were not included for training, demonstrating that the representations of reservoirs generalized well across participants.

2.3.1 Investigating temporal structure of brain data

To date, most analyses of fMRI brain data focus on understanding relatively static information (but see Introduction for further discussion). Neuronal data acquired with physiology is also most often analyzed in terms of averaged responses during certain response epochs that are believed to be behaviorally relevant. Yet, brain processes are highly dynamic and current understanding would benefit from frameworks that focus on understanding temporal processing (see [3]). Here, we employed reservoir computing to investigate and characterize temporal information in fMRI data. But the framework is sufficiently general that it can be employed with other types of brain data time series, including those from cell electrophysiology, calcium imaging, EEG, and MEG (for the use of reservoirs in other neuroscience applications, see for example [83]).

We selected two tasks from the HCP dataset to evaluate the model. The working memory task was selected as it was thought to not have a noteworthy temporal component; in the context of classification, the working memory condition was presumed to involve a relatively stable"cognitive set" (in fact, during scanning participants were informed which condition they were performing at the beginning of a block). In contrast, the theory of mind condition was expected to rely more on temporal integration of information (for every trial, participants actively attempted to discern the meaning of the vignettes so as to classify stimuli between meaningful and random).

Although classification of working memory data was a little better than ob-

tained with activation signals alone, classification did not improve with the size of the reservoir, consistent with the notion that temporal information during the block did not play a notable role in performance. In contrast, classification of data segments with meaningful social interactions (vs. random; theory of mind data) benefited from increased reservoir size. With larger memory size, accuracy improved close to 10% in some cases. As stated, the social interactions displayed in the clips build up after a few seconds and evolve throughout the block (for example, two objects "invite" a third to participate in an activity, and all three engage in it). Neuroimaging studies of theory of mind and social cognition have employed such dynamic stimuli to probe the brain correlates supporting this type of processing (for a review, see [81]). Classification between social and random clips was initially at chance levels, and increased sharply within the first 5 seconds of the clip. In future applications of the approach described here, it would be valuable to investigate how individual-level classification performance is potentially associated with behavioral performance and individual differences in social-cognitive skills (see [84]).

Whereas increases in reservoir size did not benefit working memory classification, theory of mind classification improved for theory of mind data, consistent with integration of information across time being useful for classification. Larger reservoir sizes allow signals at the same time t to interact in richer ways, too (for example, higher-dimensional signal interactions are possible). Therefore, it is possible that processing theory of mind benefited from this aspect as well (more so than working memory data), and that the correlates of theory of mind data are more distributed in the brain, and of higher inherent dimensionality (see below).

To help understand the behavior of reservoirs for classification of fMRI data, we compared the method to other temporal schemes. Although classification based on both concatenated time series data and autoregressive models performed well for working memory and theory of mind tasks, performance with reservoirs was superior to both approaches. It should be said, however, that quantitatively the improvement was relatively modest. Nevertheless, our results suggest that the nonlinear expansion of the reservoir, in addition to its temporal properties, contribute to classification performance. It should be stressed that reservoirs are straightforward to train, unlike other recurrent neural networks with fully adaptable weights. Finally, our general framework also suggests that reservoir computing provides an effective methodology to study lower-dimensional representations of the data, which may provide useful dynamic "signatures" of temporal information of fMRI data (see below).

We also investigated our proposal with naturalistic stimuli, specifically, short clips obtained from movies with either scary or funny content. Classification accuracy increased with larger reservoirs, consistent with the notion that temporal information was useful for distinguishing between the two types of clip. In the context of fMRI data which originate from hemodynamic processes with relatively slow dynamics, we suggest that the reservoir framework developed here might be particularly useful in characterizing temporal processing of naturalistic stimuli, including movies and narratives [20, 85].

2.3.2 Low-dimensional trajectories

Brain data collected with multiple techniques, including cell-activity recordings and fMRI, are often of high dimensionality. For example, calcium imaging records neuronal activation across hundreds of neurons simultaneously (for example, [86]). In fMRI, signals from tens or even hundreds of thousands of spatial locations are acquired if whole-brain imaging is considered. Even in the case where only a set of regions is of central interest, hundreds of spatial locations may be involved. Therefore, understanding lower-dimensional representations of signals is important. An important working hypothesis in cell data is that low-dimensional neural trajectories provide compact descriptions of underlying processes [3, 10].

Here, we investigated lower-dimensionality representations of reservoir states by determining classification accuracy as a function of the number of dimensions employed. For both working memory and theory of mind data, considerable reduction was attained and 12 or fewer dimensions were needed to attain classification at 95% of that obtained with the full data. Furthermore, as illustrated in Figure 2.5, even maintaining only three dimensions captured important characteristics of the ability to distinguish task conditions. More generally, we hypothesize that such low-dimensional trajectories may provide "signatures" that can be associated with tasks and/or mental states. We propose that investigating how trajectories differ across different groups of individuals (for example, low vs. high anxiety, autism vs. typically developing, etc.) is a fruitful avenue for future research. Notably, the low-dimensional trajectories captured important temporal properties of the data. For

example, for theory of mind data, trajectories were very close initially and diverged subsequently, paralleling the increase from lower to higher classification levels. These results are consistent with the idea that reservoirs provide a mechanism for the accumulation of information over time, and hence result in better accuracy in the later periods of the block.

We investigated how the dimensions with the highest contributions to distinguishing conditions were expressed in the brain by generating importance maps. In the case of working memory, several regions in lateral prefrontal cortex, parietal cortex, and anterior insula contributed to classification. These results are consistent with a large literature showing the participation of these regions in effortful cognitive functions, including working memory [87, 88]. In the case of the theory of mind task, we observed regions in the vicinity of the temporal-parietal junction and associated regions that have been implicated in theory of mind more generally, and the interpretation of social animations in particular [81]. Of interest, regions in the cuneus/pre-cuneous, which are engaged in theory of mind tasks [81, 89], were observed, too. Together, these results show that the framework developed here captures information from brain regions known to participate in the tasks investigated.

For the theory of mind data, we further compared classification accuracy obtained with the whole brain ROI partition (360 ROIs) and the lower-dimensional representations, separately, with those obtained by selecting regions from a meta-analysis across studies using social animations [81]. Intriguingly, classification with 22 targeted ROIs performed around 10% lower than obtained with the full data; it also performed more poorly than a lower-dimensional representation with only the

top/bottom 4 time series. These results raise the intriguing possibility that regions not detected in the meta-analysis carry useful information about the type of theory of mind investigated here. Therefore, to the extent that classification accuracy relies on features that are "representational," these results indicate that the correlates of theory of mind are more distributed across the brain. However, given that the present work did not determine the precise features contributing to classification, further work is needed to establish this possibility (see [4] for discussion of related issues). At the same time, we should note that lower-dimensional representations performed rather well in classifying the stimuli; therefore, representations based on a relatively low number of dimensions (for example, around 10) are feasible. For a related approach to understand the dimensionality of temporal representations in the brain, see [82].

We also studied lower-dimensional representations and temporal trajectories obtained from naturalistic movie watching. Whereas this component of our work was more exploratory, our findings revealed that the framework proposed here has the potential to be useful in these scenarios. We not only found that lower-dimensional representations could capture most of the information required for classification, but that temporal trajectories were also informative. Future work should evaluate more systematically the use of our proposal when heterogeneous stimulus sets are employed, such as the movie data investigated here.

Chapter 3: Capturing Brain Dynamics: Latent Spatiotemporal Patterns Predict Stimuli and Individual Differences

As brain data become increasingly spatiotemporal, there is a great need to develop methods that can effectively capture how information across space and time combine to form representations of mental events supporting behavior. Although fMRI data are acquired temporally, they are most often treated in a quasi-static manner. However, a fuller understanding of the mechanisms that support mental functions necessitates the characterization of dynamic properties. To address this gap we describe methods to characterize both high- and low-dimensional *trajectories* that provide "signatures" for experimental conditions (Fig. 3.1B). We investigated them at the between-participant level (in contrast to within-participant) to ascertain the generalizability of the representations created by the approach. Although our methods can be applied to any kind of fMRI data, we focused here on data acquired during movie-watching given its inherent dynamics. We address the following questions:

1. Can brain signals generated by dynamic stimuli be characterized in terms of generalizable spatiotemporal patterns? How are such patterns distributed across space and time?

Run 1		Run 2	
Clip	Length	Clip	Length
TwoMen	245	Inception	227
Bridgeville	222	SocialNet	260
Pockets	189	Oceans11	250
Overcome	65	test-retest	84
test-retest	84		
Run 3		Run 4	
Clip	Length	Clip	Length
Flower	181	HomeAlone	233
Hotel	186	Brokovich	231
Garden	205	StarWars	256
Dreary	143	test-retest	84
test-retest	84		

Table 3.1: Clips in HCP movie data. Length of each clip is indicated in seconds.

2. Understanding the dimensionality of brain representations has become an important research question in recent years ([3,10,11]). Accordingly, we sought to investigate the prediction accuracy of both high- and low-dimensional trajectories.

3. If spatiotemporal patterns capture important properties of brain dynamics, do they capture information about individual differences that are predictive of an individual's behavioral capabilities and/or personality?

3.1 Methods

We employed Human Connectome Project (HCP; [90]) movie-watching data. Participants were scanned while they watched movie excerpts (Hollywood and independent films) and other short clips (see Table 3.1 for details), which we call

"clips". Data were sampled every 1 second. All 15 clips contained some degree of social and affective content. Participants viewed clips once, except for the *test-retest clip* that was viewed 4 times. We used all movie-watching HCP data, except for 8 participants with runs missing; thus we used $N = 176$. The preprocessed HCP data included FIX-denoising, motion correction, and surface registration (details in [91,92]). We analyzed data at the region of interest (ROI) level, with one time series per ROI (average time series across spatial locations within ROI). We employed a 300-ROI cortical parcellation [93]. ROIs were also grouped based on large-scale network definitions ([94]; see Figure 3.4).

3.1.1 Long Short-Term Memory for classification of brain data

Deep neural networks (DNNs) can be used for limited temporal modeling by means of sliding windows. Recurrent NN (RNN) architectures contain feedback cycles that allow signals at the current time step to be influenced by long-term information. However, training traditional RNNs is challenging give, for example, the problem of vanishing/exploding gradients that prevent them from learning relationships beyond 5-10 time steps [95]. The gating mechanisms in Long Short-Term Memory (LSTM) networks overcome these problems [96]. Here, we employ an LSTM-based architecture to characterize spatiotemporal structure in fMRI movie data.

Brain activation at time step, x_t, was fed sequentially to an LSTM (Figure 3.1A). The cell state c_t and hidden state h_t were updated as follows:

$$\begin{bmatrix} i_t \\ f_t \\ o_t \\ \hat{c}_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} \left(\begin{bmatrix} W_{ih}, W_{ix} \\ W_{fh}, W_{fx} \\ W_{oh}, W_{ox} \\ W_{\hat{c}h}, W_{\hat{c}x} \end{bmatrix} \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix} + \begin{bmatrix} b_i \\ b_f \\ b_o \\ b_{\hat{c}} \end{bmatrix} \right), \tag{3.1}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \hat{c}_t, \tag{3.2}$$

$$h_t = o_t \odot \tanh(c_t). \tag{3.3}$$

where the input gate, i, controls the extent to which the current input is exposed to the cell state; the forget gate, f, discards prior history; and the output gate, o, controls how the cell state affects gate activations during the next time step [97]. We represent the input size (number of ROIs) by N_x and the hidden state size by N_h; other gates also have size N_h. The W matrices contain the weights. For example, W_{ix} is $N_h \times N_x$ and represents the connections between the inputs, x, and the input gate, i. The bias weights are indicated by b, and σ represents the sigmoid activation function. The operator $\odot$ represents element-wise multiplication. Each output of the LSTM, h_t, was input to a fully-connected (FC) layer used to predict the input label (here, movie clip), y_t as

$$y_t = \tanh(W_{yh} h_t + b_y). \tag{3.4}$$

The dimensionality of y_t was 15 to implement 15-way classification based on the number of clips. Note that the output of the network, based on $\max y_t$, corresponds to a time series of label predictions. We analyzed classification accuracy as a function of time to identify when clips became separable in the latent space of hidden

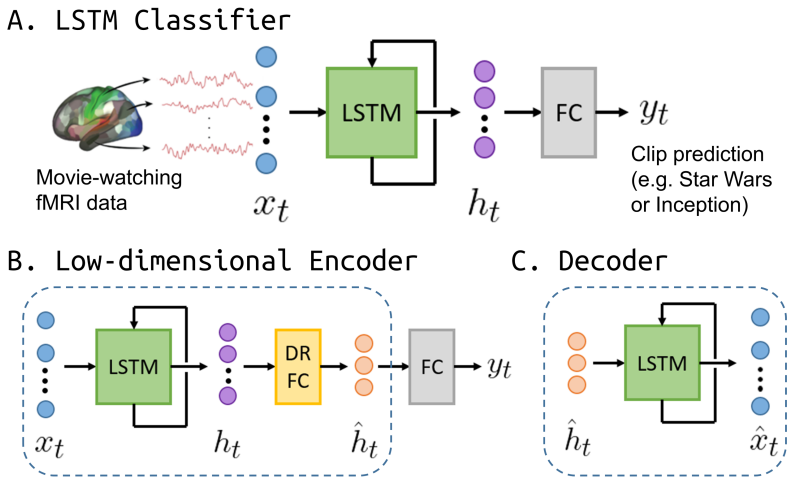

Figure 3.1: (A) The classifier consisted of Long Short-Term Memory (LSTM) units with a fully-connected (FC) dense layer for label prediction at each time step. (B) For dimensionality reduction, the LSTM outputs were first linearly projected to a lower dimensional space using a fully-connected layer (DR-FC). Classification was then performed on the low-dimensional representations, $\hat{h}_t$. (C) LSTM decoder used to reconstruct the original time series from the low-dimensional representation.

states. We refer to the LSTM together with the FC layer as the *LSTM classifier*. The classifier was trained using the Backpropagation Through Time algorithm minimizing the cross-entropy loss between the predictions and the true labels with PyTorch [98].

Data from 100 participants were used for training, and the remaining 76 participants were used for testing. To determine optimal hyperparameters for clip prediction, we employed a 10-fold cross-validation approach on the training data. In each fold, participants in the training set were not included in the validation set and vice versa.

3.1.2 LSTM-based dimensionality reduction

Unsupervised dimensionality reduction (DR) techniques such as Principal Component Analysis (PCA) project data onto a lower dimensional space such that variance is maximized. Since they do not account for class labels, representations may not be separable in the lower dimensional space. Supervised techniques such as Linear Discriminant Analysis ensure that in the projected space, data are most separable using a linear decision boundary. In most DR settings, samples are assumed to be statistically independent and thus do not account for the temporal relationships in time series data. Further, unless kernel methods are used, only linear mappings of the original high-dimensional data are possible. Here, we propose a non-linear supervised dimensionality reduction technique for temporal data.

LSTM outputs are typically high-dimensional (N_h). Several researchers have proposed "probing" into intermediate layers for increased interpretability [99, 100]. Whereas these techniques have improved understanding of how representations are learned, here we probe into LSTM hidden states to visualize dynamics. LSTM outputs, h_t, were linearly projected onto a lower dimensional space, $\hat{h}_t$, using an FC layer. We refer to this weight matrix of size $N_{\hat{h}} \times N_h (N_{\hat{h}} \ll N_h)$ as the *Dimensionality Reduction Fully-Connected (DR-FC)* layer, and to this model as the *LSTM encoder*. Since $\hat{h}_t$ is a low-dimensional representation of the history of x_t, the inputs are not treated independently, thus effectively leading to non-linear temporal dimensionality reduction. A final FC layer was used to predict labels based on $\hat{h}_t$.

3.1.3 LSTM decoder

Can low-dimensional representations obtained for classification be used to reconstruct the original data? We used an *LSTM decoder* to reconstruct x_t from $\hat{h}_t$ (Figure 3.1C). The decoder was trained separately from the LSTM encoder, and minimized the mean squared error (MSE) loss between the LSTM decoder output $\hat{x}_t$ and the input x_t. Note that the approach is different from autoencoders where latent representations are obtained such that the reconstruction loss is minimized. Here, the encoder training was independent from the decoder. To assess performance, we computed the fraction of the variance in the original data captured after reconstruction from low-dimensional representations and compared it to that captured by reconstruction from the same number of principal components.

3.1.4 Saliency maps for spatiotemporal importance

To understand the importance of each ROI at a given time step to clip prediction, we used saliency maps [101]. For each participant's clip data, the gradient of the class score with respect to the input was computed by backpropagation. To determine ROIs that were consistently important across participants, we obtained mean saliency maps for each clip by averaging across test participants.

3.1.5 Baseline models

Feed-forward network We also trained a deep feed-forward (*FF classifier*) network consisting of fully-connected layers. Each time step in the input time series was

classified independently, and thus no temporal structure was modeled. The optimal number of layers was chosen based on cross-validation following a grid search between 2 and 10 layers. The number of units in each layer was the same as the hidden state size of the LSTM.

Temporal Convolutional Network LSTMs model temporal dependency using a dynamically changing contextual window over the input time series. Are static fixed-sized windows sufficient for modeling dynamics in brain data? Thus, we employed a Temporal Convolutional Network (*TCN*; [102]), which maps an input time series of a given temporal length to an output time series of the same length (similar to LSTMs). However, convolutions are causal, such that the output at time t is based on convolutions only with inputs from time t and earlier. A fully-connected layer was used to predict clip labels based on TCN outputs, which we call *TCN Classifier*. To investigate different temporal windows, we varied the kernel width from 10 to 50 in steps of 10. By fixing the kernel height to the number of ROIs, we ensured that convolutions were only along the temporal dimension, and each kernel resulted in a 1D time series. For comparisons with LSTMs, the number of kernels was set to the optimal hidden state size of the LSTM, so that the output of the TCN was also of size N_h.

3.1.6 Predicting behavior and personality traits

LSTMs were also trained to predict behavior and personality-related scores: fluid intelligence, verbal IQ, and personality measures [65]. Using clip data as input,

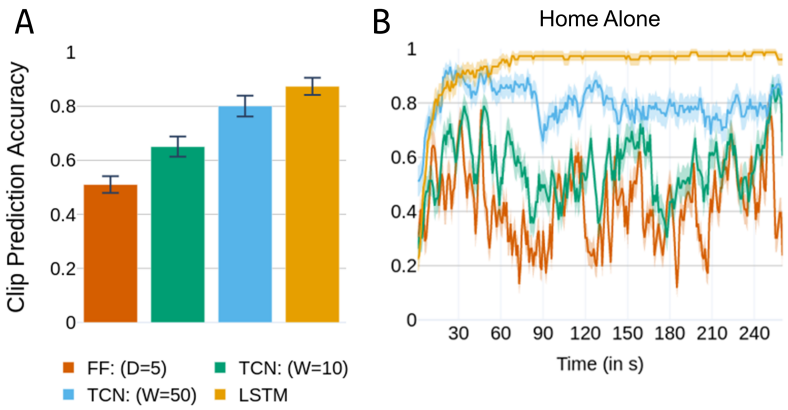

Figure 3.2: LSTMs and competing models. (A) Mean clip prediction accuracy (15-way classification) using feed-forward (FF, 5 layers) classifiers, temporal convolutional networks (TCN, kernel widths of 10 and 50), and LSTMs. (B) Since our framework predicted labels at each time step, true positive rate for each clip was determined as a function of time (illustrated for *Home Alone*). Error bars show the standard error of the mean across test participants.

an LSTM model was trained to predict behavioral scores at each time step by minimizing the MSE loss between the predictions y_t (a single unit with a continuous-valued output) and the true scores.

3.2 Results

3.2.1 Generalizability of spatiotemporal patterns in naturalistic fMRI data

To understand whether watching specific clips result in spatiotemporal brain patterns that are generalizable across participants, we employed the *LSTM classifier* to predict clip labels. The optimal hidden state size was set to $N_h = 150$ based on cross-validation (we did not find improvements using larger sizes). Classification accuracy was 87.35% for 15-way classification (Figure 3.2A). Since the outputs of

the classifier were clip predictions at each time step, we also analyzed true positive rate (TPR) as a function of time for each clip (Figure 3.2B). For most clips, TPR was poor at the beginning of the clip but increased sharply for the first 30 seconds, and then gradually reached over 90% (see Supplemental Figure B.1 for TPR of other clips). A formal evaluation of chance performance based on the null distribution obtained through permutation testing (1000 iterations; [79]) resulted in a mean null accuracy of 6.67%.

3.2.2 Is temporal information necessary for clip prediction?

We determined the extent to which distributed patterns across space (ROIs) and time contribute to clip prediction by benchmarking the LSTM classifier against competing alternatives. First, we classified based on inputs at each time step, x_t, using a feed-forward network consisting of several fully-connected layers (FF classifier). They were able to classify at no more than 51.03% accuracy (5 layers).

As FF classifiers do not capture short-term temporal relationships, we next employed temporal convolutional networks (TCN classifier). We used a kernel width (W) of 10 time steps, such that the number of parameters of the TCN classifier closely matched that of the LSTM (see Section 3.1.5 for kernel details). Classification accuracy was only 65.04%. Only with 5 times the number of parameters of the LSTM did the TCN classifier ($W = 50$) even approach LSTM performance at 83.26%. Together, the results reveal that spatiotemporal patterns are distributed across time and are most effectively captured by LSTMs capable of capturing long-

term dependencies.

Finally, if capturing temporal information and long-term dependencies are important for classification, we reasoned that performance should be affected by temporal order. To test this, we shuffled the temporal order of each clip and then trained an LSTM classifier. To preserve autocorrelation structure in fMRI data while shuffling, we used a wavestrapping approach [80]. Classification accuracy reduced drastically to 64.29%.

3.2.3 Low-dimensional trajectories as spatiotemporal signatures

We sought to characterize the intrinsic dimensionality of the data required for clip prediction. High-dimensional LSTM outputs were projected to a lower dimensional space via the DR-FC layer (Figure 3.1B). For visualization purposes, we projected to 3 dimensions. Despite the drastic dimensionality reduction of the LSTM outputs, classification accuracy was 77.30% (compared to 87.35% for 150 dimensions). In Figure 3.3A, the low-dimensional outputs are shown for each clip, which we refer to as *trajectories* (consecutive low-dimensional $\hat{h}_t$ states, $(x(t), y(t), z(t))$, describe the trajectory in "state space"). Emphasizing that high-dimensional LSTM states are important to capture temporal properties, note that the size of the hidden state layer was kept at $N_h = 150$ (Figure 3.1B); e.g., setting $N_h = 3$ reduced the classification performance drastically to less than 50%.

To visualize a notion of proximity between trajectories, we computed Euclidean distance between them as a function of time (Figure 3.3B). To compute

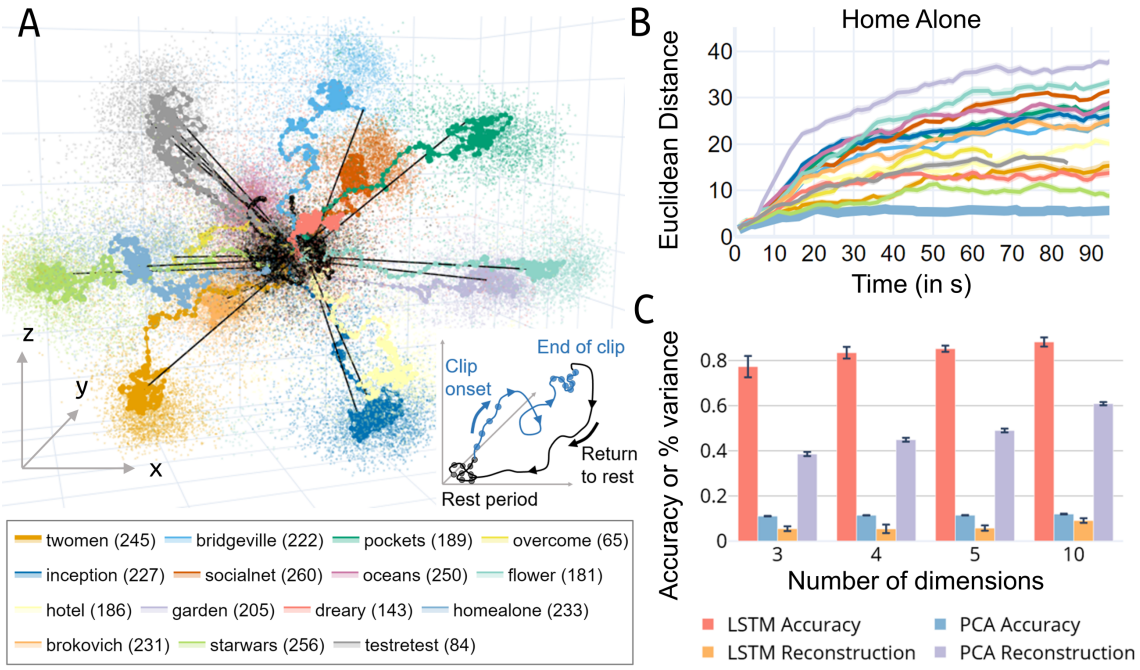

Figure 3.3: Low-dimensional trajectories. (A) Trajectories for all clips. The inset illustrates how to interpret the main result: for each clip time flows outward until the end of the clip. Solid line: mean trajectory averaged across participants; scattered points: projections for each participant at a given point in time. Trajectory associated with rest periods between clips are shown in black. (B) Distance between trajectories. The distance between the clip trajectory while watching *Home Alone* and the mean trajectory across participants for a second clip was computed. Thicker line corresponds to the distance of participants' *Home Alone* trajectories to the mean of this clip. (C) Clip prediction accuracy and fraction of variance captured after reconstruction using low-dimensional models.

the distance of clip A from B, we first computed the mean trajectory of B averaged across participants. For each participant's clip A trajectory, we computed Euclidean distance from the mean trajectory of B at every time step. Note that the proximity of a clip from itself is not zero (indicated by a thicker line), and is a measure of the consistency of participant trajectories around the clip's mean. The evolution of trajectories closely matched the temporal accuracy obtained using the original LSTM classifier. Clip trajectories were initially close-by, but slowly separated during the first minute of the clip. Plots for all other clips are shown in the supplemental material (Figure B.2).

Performance with low-dimensional encoding was surprisingly high; 3 dimensions yielded 77.30% accuracy. We further investigated low-dimensional projections with 4, 5, and 10 dimensions, at which point performance (87.23%) was very similar to that of full dimensionality. These results reveal that latent representations with as few as 10 dimensions capture essential discriminative information. However, the dimensionality-reduction fully-connected (DRFC) layer was essential in capturing this information. For reference, application of standard PCA on the input data yielded prediction accuracy at substantially lower values (Figure 3.3C). Note that the latent space uncovered by the LSTM encoder, which was successful at movie prediction even with as few as 10 dimensions, captured informational content of the fMRI time series that was substantially distinct from the fMRI signal itself. To appreciate this, consider that using the LSTM decoder (Figure 3.1C), to reconstruct the input time series, $\hat{x}_t$, from the projections of LSTM states, $\hat{h}_t$, only captured a very modest amount of signal variance (less than 10%). Again, for comparison the

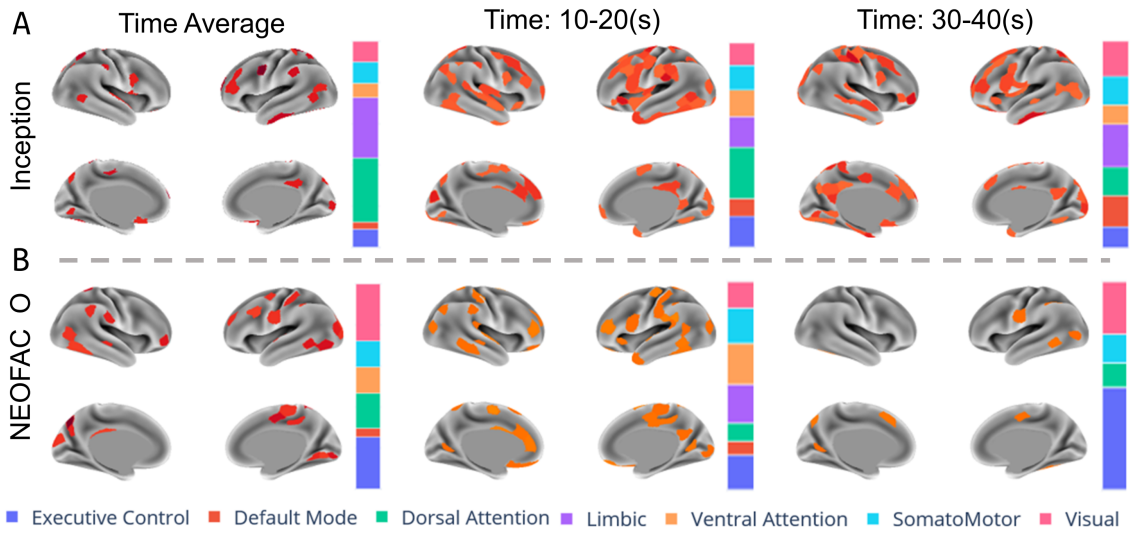

Figure 3.4: Saliency maps. (A) Top-30 most salient brain regions for predicting *Inception*. On the left, saliency was averaged across time. The right two sets of maps show mean saliency in two 10-second windows. (B) Top-30 most salient regions for predicting *openness to experience* scores (NEOFAC O). Brain maps were again based on *Inception*, the best clip for predicting the trait. The normalized contributions of 7 networks to saliency are shown alongside brain maps (colors correspond to networks and height to %).

same was done with PCA; not surprisingly, reconstructed signals recovered up to 60%.

3.2.4 Spatiotemporal saliency maps

The representations that support classification lie in an abstract space that is disconnected from the original brain activation signals. However, it is important to uncover how brain regions contribute to classification as a function of time. To do so, we used saliency maps (see Section 3.1.4). The majority of salient inputs occurred within the first 90 seconds of the clip, paralleling the increase in classification accuracy. After this period, changes at the input did not cause sizeable changes to

the class score.

To assess the contribution of different brain networks (see legend in Figure 3.4) to saliency, we averaged ROI saliency within each subnetwork to compute "network saliency", as shown in Figure 3.4A (left) for the *Inception* clip. To evaluate the evolution of saliency across time, we time-averaged ROI and network saliency for each 10-second non-overlapping window. We observed fluctuations in relative network contributions across time during the initial segment of the clip (illustrated here up to 40 s). At the level of brain regions, it is noteworthy that several ROIs with high saliency at the beginning of the clip were not captured by the time-averaged saliency map.

3.2.5 Predicting behavior and personality

In recent years, researchers have attempted to employ brain data to predict a participant's behavioral capabilities, as well as personality-based measures [30–34]. We hypothesized that spatiotemporal information captured by the LSTM architecture would provide valuable predictive information. The HCP dataset includes an extensive evaluation of each participant conducted outside the scanner. Here, we targeted available scores for fluid intelligence, verbal IQ, as well as scores associated with the NEO Five-Factor Inventory (with dimensions openness to experience (O), conscientiousness (C), extraversion (E), agreeableness (A), and neuroticism (N)).

Recent work has shown that participants with high scores along particular behavior/personality dimensions have similar brain responses to naturalistic stimuli; in

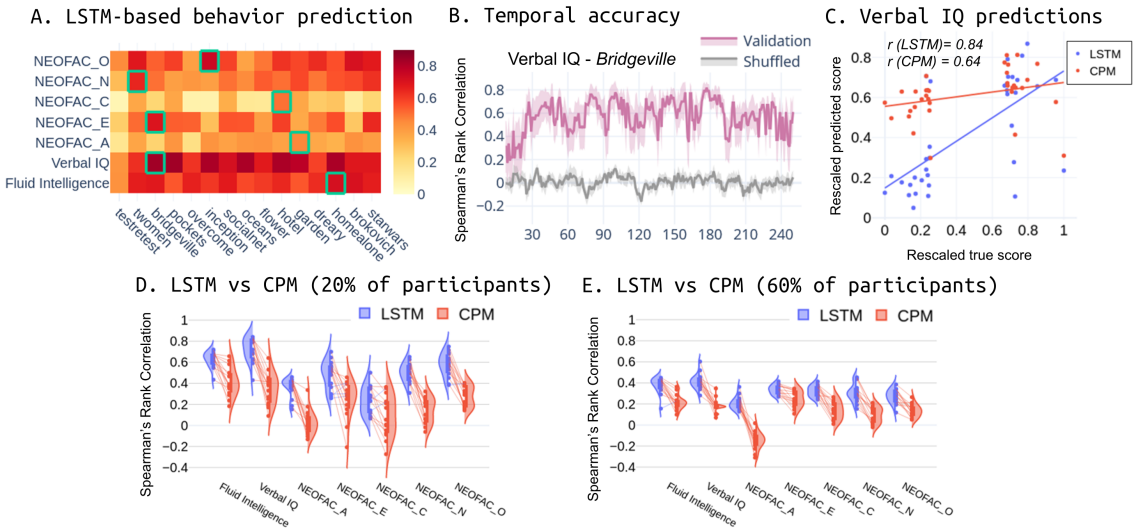

Figure 3.5: Prediction of behavior and personality. (A) Behavior prediction accuracy using LSTMs (best clip with green box). (B) Temporal accuracy for Verbal IQ while watching *Bridgeville* (best clip) along with mean null accuracy obtained by randomly shuffling behavioral measures across participants. (C) Correlation (r) between predicted and true verbal IQ scores (rescaled between 0 and 1) while watching *Bridgeville*. (D) Comparison of LSTM prediction accuracy with connectome-based predictive modeling (CPM) using top/bottom 10% of the scorers. Connecting lines indicate accuracy for the same clip. (E) Same as (D) but using top/bottom 30% of the scorers.

contrast, low scorers had more idiosyncratic responses that were less similar to both high scorers and other low scorers [103]. Accordingly, here we focused on data from participants with scores at the top and bottom 10% of scores. Based on this criterion, models were trained for each behavior/personality (approx. 40 participants) separately, and generalizability was assessed using a 10-fold nested cross-validation instead of a held-out test set due to limited data. We hypothesized that participants with particular traits would generate brain states that emphasize their personality during brief clip segments, given specific stimuli during those segments (e.g., negative or positive scenes). Accordingly, we anticipated that prediction accuracy would vary across time. We thus determined optimal hyperparameters as well as the time step when prediction accuracy was highest, t_{best}, for each clip based on the inner cross-validation. We report accuracy at t_{best} based on the outer cross-validation (Figure 3.5).

Existing techniques for predicting behavior and personality traits use what is termed *functional connectivity* as inputs to regression models [30, 32]. Functional connectivity refers to the correlation between brain regions, typically the Pearson correlation between their time series. We compared our approach to connectome-based predictive modeling (CPM; [104]), possibly the state-of-the-art in this regard. In this approach, a functional connectivity matrix is initially formed for each clip based on all ROIs. Subsequently, the entries in the matrix (or edges) that correlate with the behavioral measure beyond a set threshold (here, 0.2) are retained and a linear model is fit to predict behavioral or personality scores based on these edge weights (i.e., the functional connectivity). Unlike LSTMs, predictions are made

based on the entire clip, and not at each time step.

We used all clips for training (rather than training a separate model based on each clip) to promote learning representations that are not idiosyncratic to a particular clip, and thus generalizable across clips. Nevertheless, model performance, measured as Spearman's rank correlation between predicted and measured scores, was computed for each clip to understand how behavior/traits were captured better by certain stimuli. LSTM accuracy varied across behavior/personality measures and clips, but was consistently and robustly higher than CPM (Figure 3.5A and D). Note that chance performance for each measure based on permutation testing (1000 iterations; [104]) resulted in correlations of less than 0.1. We illustrate the correlation between predicted and true scores in the case of verbal IQ (Figure 3.5C). For other participant cutoffs, see Figure B.3. To ascertain that the improved performance of LSTM was not driven by the selection of the 10% cutoff, we increased the number of participants to include the top/bottom 30% of the scorers. Again, LSTMs performed considerably better than CPM, although prediction accuracy decreased for both models (Figure 3.5E). Comparisons between LSTMs and CPM for other participant cutoffs are shown in Figure B.4. To illustrate the temporal dimension of performance, we used the *Bridgeville* clip to predict Verbal IQ (Figure 3.5B). As observed with most clips (not shown), accuracy increased at the beginning of the clip but continued to fluctuate. The results support the idea that behavior and personality traits are best captured during specific clip segments.

Finally, we sought to determine the importance of brain regions and brain networks to prediction via saliency maps (Figure 3.4B). For prediction of openness to

experience scores based on the *Inception* clip, the time-averaged saliency map across the entire clip (left) did not capture several salient regions at other time windows, again illustrating the temporal dimension of saliency maps. Finally, it is noteworthy that brain regions involved in classification of the *Inception* clip (Figure 3.4A) were quite different from those involved in predicting the openness personality trait.

Chapter 4: Comparing Functional Connectivity Matrices: A Geometry-Aware Approach applied to Participant Identification

Understanding the correlation structure associated with multiple brain measurements (acquired across multiple sensors or spatial locations) is a central goal in neuroscience, as it informs about potential "functional groupings" and network structure [2, 105]. The correlation structure can be conveniently captured in a matrix format that captures the relationships among a set of brain measurements. For example, in the case of fMRI, each entry of the matrix might contain an estimate of the *functional connectivity* (FC) between regions i and j, typically computed as the correlation between the time series data of the two regions in question.

In recent years, the FC matrix has become an important component of many types of investigation focusing on network-level properties of the brain, particularly in fMRI. For example, it has been used to cluster brain states [40], characterize dynamic functional states [106], perform participant identification [41], and understand how tasks reconfigure brain networks [107]. In these applications, some notion of proximity or similarity of FC matrices is employed (Figure 4.1A). How should similarity be gauged? An intuitive approach is to "unroll" the FC matrix into a

vector and compute the Pearson correlation between the matrices themselves. Thus if, say, two brain states captured by FC matrices are similar (for example, during two similar perceptual conditions), their matrices would be (relatively) highly correlated. Indeed, the correlation approach has yielded impressive results, such as successfully identifying a participant out of a large group of participants based on FC matrix similarity, a process dubbed fingerprinting [41, 108, 109]. Related approaches include computing the Euclidean (L^2) distance between the vectorized matrices [110], or using the so-called Manhattan (L^1) distance [40].

FC matrices computed by Pearson correlating time series data are objects that lie on a non-linear surface (technically known as a manifold) called the *positive semidefinite cone*: their geometry is non-Euclidean. Accordingly, distances between Pearson FC matrices must be measured along the surface of the cone. In addition, FC matrices are often high dimensional, and the proximity measure adopted is critical since noisy dimensions can contribute substantially to the measure [111].

In this chapter, we characterize the advantages of using a *geodesic* proximity measure between FC matrices. We apply the approach to the problem of participant identification: Given resting-state or task data, is it possible to determine a participant from her FC matrix [41]? We show that using the geodesic distance, a non-Euclidean distance metric that considers the manifold on which the data lies, improves participant identification compared to a similarity measure based on Pearson correlation (Figure 4.1C). The improvement is shown to be non-trivial and consistent across resting-state and task conditions.

We also investigate how distances between high-dimensional FC matrices can

be effectively visualized in low-dimensional spaces. Such visualization reflected identification accuracy based on the full-dimensional data, and thus retained important distance information. We suggest that visualization in lower dimensions aids in understanding the geometry of task FC structure in relation to resting-state FC.

4.1 Methods

4.1.1 Human Connectome Project Data

We utilized data from $N = 100$ unrelated participants from the Human Connectome Project (HCP) of the 1200-participant release [112]. Data from resting-state and seven tasks were employed: emotion processing (EM), gambling (GB), language (LG), motor (MT), relational processing (RL), social cognition (SO), and working memory (WM). Throughout the paper, we refer to resting-state plus the tasks as *conditions*. For a description of the tasks and scan parameters, see [113]. Data were collected with a repetition time (TR) of 720 ms.

During each run, stimuli were presented in separate blocks often interleaved with fixation blocks. Some task runs also contained cues. To retain only task-related segments of the run, extraneous segments were trimmed. To account for hemodynamic lag, the first four TRs of the block were not used, and the first four TRs following the end of the block were used [114]. Emotional processing, working memory, and motor tasks contained 3-second cues at block onset. Accordingly, to account for the cue response and the hemodynamic lag, data from 12 seconds after the cue onset to 3 seconds after the end of the block were used. Time course length for

each condition before and after trimming is provided in Table 4.1. Note that trimming the fixation periods is important in characterizing participant identification from task data, because fixation periods behave much like "mini resting periods" that can potentially provide information regarding the participant. Analysis of data without trimming is included in supplemental material (Section C.1).

Condition	REST	EM	GB	LG	MT	RL	SO	WM
Frames per full run	1200	176	253	316	284	232	274	405
Frames per trimmed run	1200	141	156	295-305	170	138	160	312

Table 4.1: Number of frames per run (in samples) before and after trimming fixation periods.

4.1.2 Preprocessing

Task data were part of the "minimally preprocessed" release, which included gradient unwarping, fieldmap-based EPI distortion correction, brain-boundary-based registration of EPI to structural T1-weighted scan, non-linear registration, and intensity normalization [115]. Cortical data were mapped to a surface representation and utilized here. In addition, we regressed out 12 motion-related variables (6 translation parameters and their derivatives) and low frequency signal changes using the 3dDeconvolve program of the AFNI package [67] with the `ortvec` and `polort` options (the latter removed linear, quadratic, and cubic trends over the duration of individual runs. Resting-state also followed the so-called minimal preprocessing pipeline, in addition to denoising using ICA-FIX [116] and regressing out 12 motion-related variables, as provided with the data distribution. Cortical data

were mapped to a surface representation. Preprocessing included minimal temporal filtering that essentially removed linear trends in the data. The ICA-FIX procedure removed "bad" components such as high frequency noise from the data. No further preprocessing was performed for resting-state data in the main text. In particular, band pass filtering is not included in HCP's preprocessing because they believe it can potentially eliminate relevant information in resting-state data [117].

For the results in the main text, the global mean was not regressed from the data. In the supplemental material (Section C.2), we repeated some analyses on resting-state data that included global signal regression as part of the preprocessing pipeline. Although there is no consensus in the field whether or not the global mean should be eliminated, some work has reported that removal strengthens the association between resting-state functional connectivity and behavior [118, 119].

4.1.3 Regions of interest and organization into subnetworks

For simplicity, we focused on cortical regions of interest (ROIs) only. We used the local-global Schaefer cortical parcellations that divide the cortex into 300 ROIs [120] (throughout the text, we refer to it as "whole-cortex"). A summary ROI-level time series was obtained by averaging signals within the region. We then used the Yeo 7-network parcellation to group the ROIs into 7 subnetworks known as `visual, somatomotor, dorsal attention, ventral attention, limbic, frontoparietal, and default mode` [94]. The number of ROIs within each of the subnetworks is provided in Table 4.2. The ROIs and the

grouping into 7 networks is shown in Figure C.4. Some of the effects of varying the number of ROIs are described in the supplemental material (Section C.3).

Subnetwork	Number of ROIs
Visual	47
SomatoMotor	57
Dorsal Attention	34
Ventral Attention	34
Limbic	20
FrontoParietal	40
Default	68

Table 4.2: Number of ROIs in each subnetwork. We used local-global Schaefer cortical parcellations that divide the cortex into 300 ROIs [120].

4.1.4 Functional connectivity

Functional connectivity was computed by Pearson correlating time series data between every pair of ROIs, resulting in 300×300 FC matrices. A symmetric matrix S that satisfies $y'Sy \geq 0$ (where y' is the transpose of y) for any non-zero vector y is said to be positive semidefinite and has eigenvalues greater than or equal to zero. Though it is well known that covariance matrices are positive semidefinite [121], we illustrate the proof here. After normalizing the time series of each ROI to unit variance, let $x_t = (x_{t,1}, x_{t,2}, \ldots, x_{t,300})$ be the vector of activations of all ROIs at time t for $t = 1, 2, \ldots, T$. If we denote the mean across time as $\bar{x}$, the covariance matrix is given by

$$Q = \frac{1}{T} \sum_{t=1}^{T} (x_t - \bar{x})(x_t - \bar{x})'. \tag{4.1}$$

Note that the (i, j) entry of Q is simply the Pearson correlation coefficient between the time series of regions i and j. For any non-zero vector y of dimension 300,

$$\begin{aligned} y'Qy &= y' \left(\frac{1}{T} \sum_{t=1}^{T} (x_t - \bar{x})(x_t - \bar{x})' \right) y \\ &= \frac{1}{T} \sum_{t=1}^{T} y'(x_t - \bar{x})(x_t - \bar{x})'y \\ &= \frac{1}{T} \sum_{t=1}^{T} ((x_t - \bar{x})'y)^2 \geq 0. \end{aligned} \tag{4.2}$$

Thus, covariance matrices are positive semidefinite.

If Q_1 and Q_2 are two FC matrices, it can be easily shown following the steps above that $\alpha Q_1 + \beta Q_2$ is also positive semidefinite for $\alpha, \beta > 0$. Thus, the set of all positive semidefinite matrices lie on a cone referred to as the *positive semidefinite cone* [121].

4.1.5 Geometry of functional connectivity matrices

Pearson correlation is often used to characterize the similarity of FC matrices. However, as correlation matrices lie on a non-linear space, a natural approach is to compute *geodesic distances* between FC matrices to quantify their distance. The geodesic distance between two points on the positive semidefinite cone, and thus between two FC matrices Q_1 and Q_2, is the shortest path between them along the manifold [122]. There exists only one geodesic path joining two such points.

For two functional connectivity matrices, their geodesic distance can be computed as proposed in [122]:

$$d_G(Q_1, Q_2) = \sqrt{\text{trace}(\log^2(Q_1^{-\frac{1}{2}} Q_2 Q_1^{-\frac{1}{2}}))}, \tag{4.3}$$

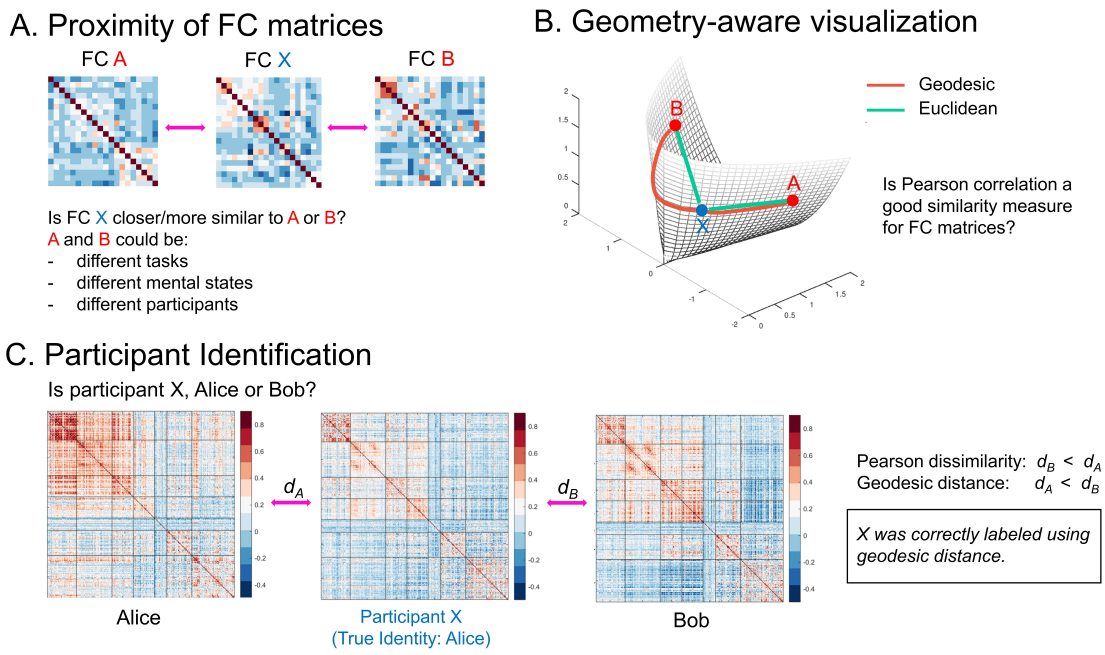

Figure 4.1: Functional connectivity matrices and their underlying geometry. (A) Similarity of functional connectivity (FC) matrices. Is the FC matrix X more similar to A or B? This question arises when the goal is to determine the task being being performed, the mental state, or the participant. (B) Illustration of geodesic distance (red) and Euclidean distance (green) on the so-called positive semidefinite cone. The geodesic and Euclidean distances between two points can differ substantially. (C) Is X, Alice or Bob? Equivalently, is the FC X more similar to that of Alice or Bob? Identification involves mapping an unknown participant's data to one of the participants in the database (only two in this case). In this example, X is correctly labeled as Alice using geodesic distance, but incorrectly labeled as Bob using Pearson dissimilarity.

where the matrix log operator is used here. Note that this definition assumes that the matrix Q_1 is invertible; when this was not the case the identity matrix, I, was added as a perturbation matrix to both Q_1 and Q_2 to ensure that all eigenvalues were greater than 0 (see Section C.4). For matrices Q_1 and Q_2 of size $n \times n$ (here, $n = 300$ ROIs), if $Q = Q_1^{-\frac{1}{2}} Q_2 Q_1^{-\frac{1}{2}}$, and λ_i for $i = 1$ to n are the n eigenvalues ≥ 0 of Q, the geodesic distance is simply (see `https://github.com/makto-toruk/FC_geodesic`

for code)

$$d_G(Q_1, Q_2) = \sqrt{\sum_{i=1}^{n} (\log(\lambda_i))^2}. \tag{4.4}$$

From (4.4), it is clear that $d_G \geq 0$. In addition, $d_G = 0$ implies $\lambda_i = 1$ (i.e, $Q_1 = Q_2$), and vice versa. To verify that the geodesic distance is symmetric, note that $d_G(Q_1, Q_2) = d_G(Q, I)$ (using Eq. 4.3). By the property of the log operator, $d_G(Q, I) = d_G(I, Q)$ since $\log^2(Q^{-1}) = \log^2(Q)$. We refer the interested reader to [123] for a proof of the triangular inequality for 2×2 matrices. Thus, the geodesic distance applied to matrices meets the criteria of a *metric*.

If q_1 and q_2 are vectors obtained by stacking the columns of Q_1 and Q_2, respectively, Pearson dissimilarity between the two matrices is defined as

$$d_P(Q_1, Q_2) = \frac{1 - \text{corr}(q_1, q_2)}{2}, \tag{4.5}$$

where the corr function is the Pearson correlation coefficient. Pearson dissimilarity ranges between 0 and 1 and is *not* a formal metric because it does not satisfy the triangular inequality [124]. The units for geodesic distance and Pearson dissimilarity are arbitrary and thus not comparable across these measures.

4.1.6 Participant identification

Identification involves mapping an unknown participant's data to one of the participants in the database. Since each task in the HCP data contains 2 runs for every participant, we used one run as training data (that is, to form the database) and the other run for testing. Identification was performed on each condition (resting-state or task) separately.

Participant identification is equivalent to N-class classification where the objective is to label an individual's FC matrix in the test data to one of the N participants in the training data. To do so, we used a 1-Nearest Neighbor approach [41]: An FC matrix in the test data is labeled with the participant identity of the FC that is most similar to it in the training data. Suppose Q_x^{test} is an unknown participant's FC matrix. Then

$$\text{label}(x) = \arg\min_{i=1}^{N} d(Q_i^{\text{train}}, Q_x^{\text{test}}), \tag{4.6}$$

where Q_i^{train} is the ith participant's FC matrix in the training data and $d(\cdot, \cdot)$ is a distance or similarity measure. Here we compare the use of a geodesic distance metric to a Pearson dissimilarity measure.

4.1.6.1 Identification accuracy

Participant identification was performed using the first run as training data and the second run as testing data. For the N participants in the testing data, accuracy was defined as

$$\text{Accuracy} = \frac{\text{Number of correctly labeled participants}}{\text{Total number of participants}}. \tag{4.7}$$

Then, the roles of the training and testing data were reversed and accuracy was computed again. The reported identification accuracy was the mean of the two accuracy values.

4.1.7 Bootstrapping

For participant identification statistics, one must confront the non-independence between participants in the sample. Consider the following case. If two participants' FC matrices Q_A and Q_B are close to each other, B might be mislabeled as A. However, if A was not in the training database, it is conceivable that B would have been labeled correctly. Therefore, the *entire group* must be considered as the unit of interest; it is the group that determines if identification performance will be poor or good. In our study, we used data from $N = 100$ participants in the age range of $22 - 35$ years, but demographic factors such as age and mental health status can potentially play an important role in identification performance.

A convenient procedure to assess variability in identification performance is to use bootstrap resamples, with each resample comprising random draws with replacement of the urn containing the group of participants. Thus, a bootstrap resample is a proxy for a group of participants, and variability can be quantified by resampling it a large number of times.

More precisely, suppose a dataset of size N for a run is denoted by $\mathcal{D}$. Let $0 \leq f_P(\mathcal{D}) \leq 1$ and $0 \leq f_G(\mathcal{D}) \leq 1$ be the participant identification accuracy obtained using Pearson dissimilarity and the geodesic distance, respectively. Let $\mathcal{R}_j$ be a dataset also of size N obtained by resampling $\mathcal{D}$, with replacement, N times. Thus, $\mathcal{R}_j$ is a bootstrap resample of $\mathcal{D}$ and may contain duplicate entries. The accuracy difference on this bootstrap resample is given by $\delta(\mathcal{R}_j) = f_G(\mathcal{R}_j) - f_P(\mathcal{R}_j)$. Such difference score is computed for $M = 1000$ bootstrap resamples $\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_M$ and

the *mean* difference score, $\bar{\delta}$, is computed. This process (based on M resamples) provides exactly one mean difference score. The question of interest is as follows: How are *mean* difference scores distributed? Note that this parallels the question of the distribution of the sample mean in the setting of the standard Central Limit Theorem. In our case, the distribution of *mean difference scores* is of interest. Since the object of interest is the *mean* difference score, the procedure to determine a specific $\bar{\delta}$ is repeated $B = 1000$ times, resulting in $\{\bar{\delta_1}, \bar{\delta_2}, \ldots, \bar{\delta_B}\}$ (that is, B mean differences). Although the number of resamples, $M \times B$, is large, the distance matrix of size $N \times N$ (between each subject's test-FC to all subjects' train-FC) is calculated only once making the bootstrapping procedure computationally feasible.

Reported p-values were computed as follows. Because accuracy differences are percentages, we initially applied a standard Fischer-z transformation to $\{\bar{\delta_1}, \bar{\delta_2}, \ldots, \bar{\delta_B}\}$ so that their distribution would be approximately normal. To test the null hypothesis $H_0 : \bar{\delta} = 0$, a one-sample t-test was then used.

4.1.7.1 Evaluating shorter data segments

To understand the effect of the length (or the number of frames) of the run, we truncated runs to smaller segments. For a particular segment length, 50 segments were obtained each of which had a unique, randomly-chosen starting point in the run. The objective was to pick several segments of the same length without favoring those that started at the beginning of the scan. For each segment, 1000 bootstrap iterations were used to obtain a mean accuracy score.

4.1.8 Multidimensional scaling

Naturally, visualizing distances between FC matrices is not straightforward given their high dimensionality. Here, we used *non-metric multidimensional scaling* to visualize distances in three dimensions [125]. Whereas standard multidimensional scaling computes the Euclidean distance between the high-dimensional vectors of interest, non-metric multidimensional scaling takes as input any *dissimilary matrix* of the form

$$D = \begin{bmatrix} d_{1,1} & d_{1,2} & \dots & d_{1,200} \\ d_{2,1} & d_{2,2} & \dots & d_{2,200} \\ \vdots & \vdots & \ddots & \vdots \\ d_{200,1} & d_{200,2} & \dots & d_{200,200} \end{bmatrix} \tag{4.8}$$

where $d_{i,j}$ is the "dissimilarity" between the FC matrices i and j (the dimensionality of the matrix is 200 since we consider a test-FC and a train-FC for each of the $N = 100$ participants). Here, either geodesic distance or Pearson dissimilarity was used. Given D, non-metric multidimensional scaling finds a set of $\mathbb{R}^3$ vectors such that the Euclidean distance between these vectors preserves, to the extent possible, the high-dimensional distances:

$$\hat{d_{i,j}} = ||x_i - x_j||_2^2 \approx d_{i,j} \tag{4.9}$$

where the vectors x are low dimensional. Thus, if $d_{i,j} = d(Q_i, Q_j)$ is the distance between two FC matrices Q_i and Q_j, and $\hat{d_{i,j}}$ is the distance in the lower-dimensional representation, the output (set of points) is produced by minimizing the *stress* func-

tion:

$$S = \sqrt{\frac{\sum_{i<j}(d_{i,j} - \hat{d_{i,j}})^2}{\sum_{i<j} d_{i,j}^2}}. \tag{4.10}$$

The optimal distances, $\hat{d_{i,j}}$, are obtained using a gradient descent approach that minimizes the stress. The MATLAB 2018a [126] implementation of *mdscale* with 1000 gradient descent iterations was used. Multidimensional scaling produces low dimensional representatives, x's, for high dimensional FCs such that the Euclidean distances between x's approximate the measured relationships (Pearson dissimilarity or geodesic distance) between their high-dimensional counterparts. Given that the two measures have arbitrary units, so do their estimates in low dimensions.

Note that the objective of using non-metric multidimensional scaling was to represent in a more intuitive manner the relationships between high-dimensional functional connectivity matrices. Thus, points in the lower-dimensional representation no longer lie on the positive semidefinite cone and *closeness* should be interpreted in the Euclidean sense (two points are close if their Euclidean distance is small). The visualizations, approximate as they are, are only provided to aid understanding, and are not part of the procedure to determine identification accuracy.

4.1.9 Note on p-values

As discussed by many others recently, we do not view "statistical significance" dichotomous thresholds (for example, $p < 0.05$) as the ultimate criterion in deciding whether a result is "real" or not [127, 128]. In any case, understanding variability and the unlikeliness of a result provides some information. Given that we compare

geodesic distance to Pearson dissimilarity across conditions and other parameters, some form of correction for multiple comparisons is opportune. Accordingly, we provide the uncorrected p-value as well as the Bonferroni-corrected α level (which we call the "reference α") so that the reader can further gauge the "strength" of the finding. Again, we do not advocate using the Bonferroni-corrected α in a dichotomous fashion, but provide it as an additional "reference" point for the reader.

4.2 Results

4.2.1 Motivation behind geodesic distance

We motivate the geodesic distance with simple examples from the space of 2×2 FC matrices. Since FC matrices are symmetric and positive semidefinite, they take the form

$$Q = \begin{bmatrix} x & y \\ y & z \end{bmatrix},$$

and satisfy $x \geq 0, y \geq 0$ and $xy - z^2 \geq 0$. Since the matrices have only three unique entries, all points that satisfy these equations can be plotted in three dimensions in Euclidean space, and form a *positive semidefinite cone* (Figure 4.1B).

In the first example, we considered three points on the cone (i.e., three 2×2 FC matrices) 'a', 'b' and 'c' such that 'b' and 'c' are equidistant from 'a' in terms of the Euclidean distance (Figure 4.2A). If a tangent surface to the cone is drawn at 'a', the point 'c' is much closer to the tangent surface than 'b'. Thus, the geodesic

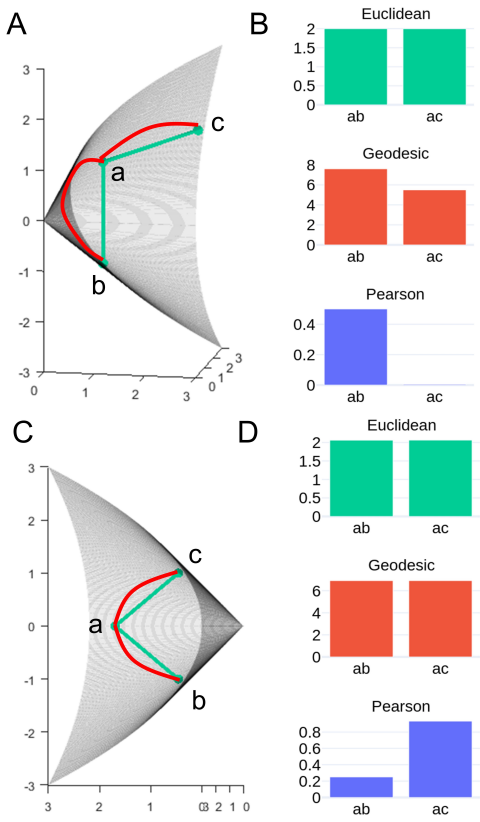

Figure 4.2: Motivating functional connectivity geometry. (A) Identical Euclidean distance does not imply identical geodesic distance. (C) Identical geodesic distance can yield very different Pearson dissimilarity. (B, D) Comparison of distances/dissimilarity *ab* and *ac* in (A) and (C), respectively. Distances/dissimilarity cannot be compared across measures because their units are arbitrary.

distance between '*a*' and '*b*' is larger than that between '*a*' and '*c*' (Figure 4.2B). In this case, Pearson dissimilarity is capable of distinguishing the two distances.

To motivate why Pearson dissimilarity is problematic, consider that the Pearson correlation between two vectors is equivalent to the cosine of the angle between them after they have been "centered" individually (that is, the mean of each vector is subtracted from it) and normalized. Indeed, the computation of Pearson correlation eliminates the contribution of the signal mean, as can be readily seen in the

following equation:

$$\mathrm{corr}(x, y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}, \tag{4.11}$$

where x and y are vectors. For FC matrices, such centering which is implicit in Pearson correlation alters the eigenvalues and the positive semidefiniteness of the matrix. Since the eigenvalues are the basis for computing geodesic distances, we see that Pearson correlation in fact distorts the evaluation of similarity between connectivity matrices (relative to what is estimated with the geodesic distance). However, while estimating an individual's FC matrix, mean centering does not affect positive semidefiniteness, as shown in Eq. (4.2).

In a second illustrative example (Figure 4.2C), we consider three points 'a', 'b' and 'c' on the cone such that 'b' and 'c' are symmetrically on either side of 'a'. By symmetry, 'a' is equidistant from 'b' and 'c' in terms of both the Euclidean distance and geodesic distance. However, Pearson dissimilarity between the two sets of points can be quite distinct. Suppose O is the origin and $\angle aOb = \angle aOc$ (where $\angle$ is the angle subtended between 'a' and 'b'). Since Pearson correlation mean centers the vectors 'a' and 'b', the correlation is related to mean-centered vector angles that can be quite different from the original ones (Figure 4.2d). In other words, if '$\underline{a}$', '$\underline{b}$', and '$\underline{c}$' are vectors obtained by centering 'a', 'b' and 'c', in most cases $\angle \underline{a}O\underline{b} \neq \angle \underline{a}O\underline{c}$. The upshot is that measures of similarity based on Pearson correlation do not correspond to actual distances between functional connectivity matrices.

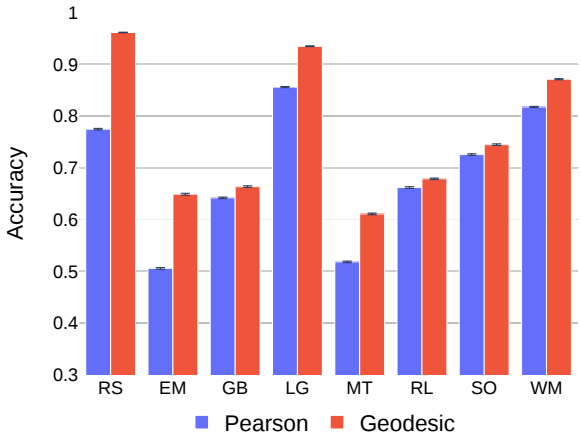

Figure 4.3: Participant identification for the eight conditions using the geodesic distance and Pearson dissimilarity. Training and testing data were from the same condition. Accuracy improved using the geodesic distance on each condition. Error bars indicate standard error of the mean across bootstrap iterations. Abbreviations: EM, emotion processing; GB, gambling; LG, language; MT, motor; RL, relational processing; RS, resting-state; SO, social cognition; WM, working memory.

4.2.2 Geodesic distance and participant identification

Participant identification ($N = 100$) was performed on each condition (resting-state and tasks) using two measures: geodesic distance and Pearson dissimilarity (Methods 4.1.6). FC matrices obtained from one run were used as training data and matrices from the second run as testing data. Identification accuracy for each condition is shown in Figure 4.3 (accuracy based on chance would be 1%).

To assess the robustness of the results and for statistical comparisons between the two measures, identification was performed on bootstrap resamples. For each bootstrap resample, the difference between accuracy using geodesic distance and Pearson dissimilarity was computed. A one-sample two-tailed t-test was then

used to assess the null hypothesis that the difference distribution had zero mean (Methods 4.1.7). For each condition, using the geodesic distance improved identification accuracy over Pearson dissimilarity ($p < 10^{-6}$ for all conditions; reference $\alpha = 0.05/8 = 0.00625$ given 8 conditions; Figure C.6). The mean improvement using geodesic distance was around 8%, ranging from 2% (*relational processing*) to as much as 19% (*resting-state*). For *resting-state* and the *language* conditions, the accuracy obtained using the geodesic distance was very hight and close to 95%.

Finn et al. [41] reported a mean accuracy of 93.65% on *resting-state* data using Pearson dissimilarity, which is considerably higher than the 77.5% we obtained. Given that in the HCP dataset four runs of *resting-state* data are available per participant (collected over separate days), they averaged the FC matrices obtained obtained during the same day into a single FC matrix. By including this averaging procedure, we replicated their findings more closely and obtained an accuracy of 91% using Pearson dissimilarity. Using geodesic distance, accuracy increased to 98%. However, since conditions other than *resting-state* contained only two runs, we did not use the averaging procedure on the four runs of resting-state data in the remainder of our work.

4.2.3 Low-dimensional visualization of functional connectivity matrices

Since FC matrices are high dimensional, multidimensional scaling was used to visualize the distances between them in three dimensions (Figure 4.4). The goal

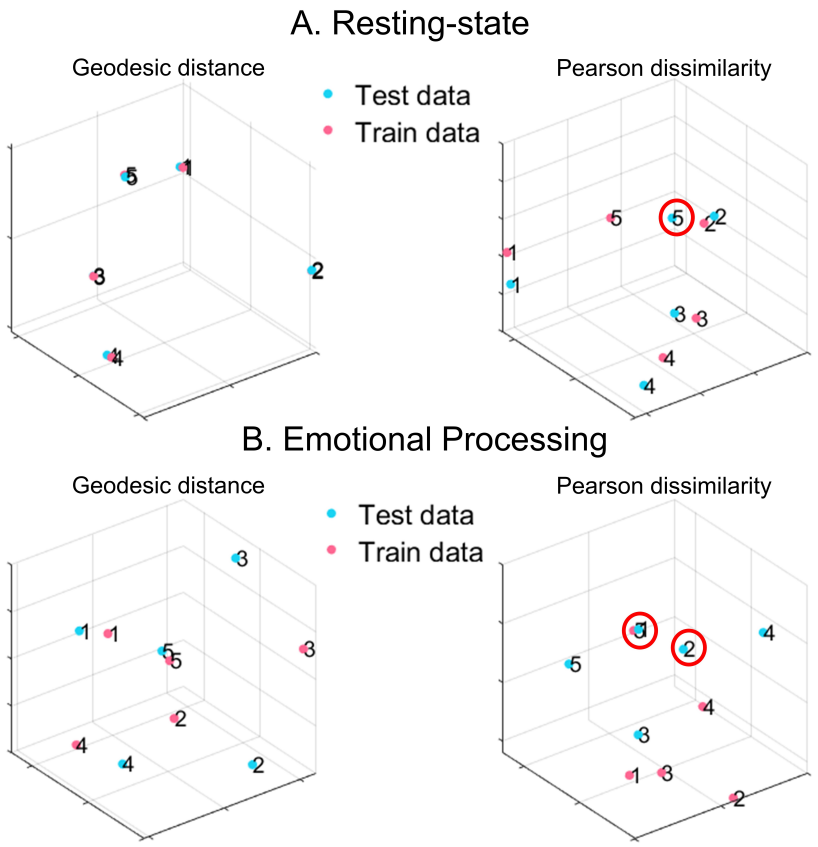

Figure 4.4: Visualization of geodesic distance and Pearson dissimilarity. Distance/similarity between high-dimensional functional connectivity matrices (300 × 300) was visualized in three dimensions using non-metric multi-dimensional scaling. Training data (blue) and testing data (pink) were selected from five random participants (numbers 1-5). Mislabeled participants are encircled in red. (A) *Resting-state*. (B) *Emotional processing* task. For *resting-state*, within-participant geodesic distances were very small relative to between-participant distances in the lower-dimensional representation (when numbers labeling the participants overlapped, only one of them is visible). Online figures are available [129].

of using multidimensional scaling was to represent in a more intuitive manner the relationships between high-dimensional FC matrices. Accordingly, points in the lower-dimensional representation should be interpreted in the Euclidean sense (two points are close if their Euclidean distance is small). But note that the visualizations are approximate only, and provided to aid understanding (they are not part of the procedure to determine identification accuracy).

Within- and between-participant distances estimated in three dimensions were indicative of varying identification accuracy (obtained using high dimensional FC matrices) across conditions. For *resting-state*, FC matrices within-participant geodesic distances between training and testing were very small, whereas distances between different participants were considerably larger, consistent with the high identification accuracy. Visualization of Pearson dissimilarity revealed similar characteristics, but the ratios of within- to between-participant distances were not as large. In fact, using Pearson dissimilarity resulted in participant 5 being mislabeled as participant 2, for example.

For the *emotional processing* task, within-participant distances were *not* much smaller than between-participant distances even for the geodesic distance consistent with the lower accuracy on this task. However, all participants in the randomly chosen subset were still labeled correctly. Using Pearson dissimilarity, two participants were mislabeled. In general, using the geodesic distance resulted in more favorable ratios of within- to between-participant FC distances.

4.2.4 Identification accuracy and time course length: resting-state data

Since the length of the time course plays a key role in the quality of the estimate of the FC matrix [130, 131], we sought to characterize its effect on participant identification. Because *resting-state* data had the longest time course (1200 TRs), shorter segments varying from 100 to 1100 TRs (in steps of 100) were extracted.

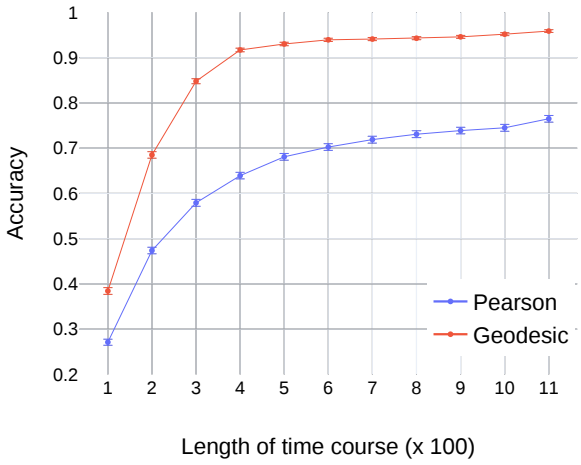

Figure 4.5: Participant identification accuracy as a function of segment length for *resting-state* data. Accuracy using geodesic distance exceeded Pearson dissimilarity at each segment length (see text). Error bars indicate standard error of the mean across bootstrap iterations.

Accuracy improved with length for both measures (Figure 4.5). Accuracy using the geodesic distance was higher than Pearson dissimilarity for segment lengths greater than 200 TRs ($p < 10^{-4}$; reference $\alpha = 0.05/11 = 0.0045$ given 11 segment lengths; Figure C.7). For segment length of 100 TRs, accuracy using geodesic distance was still higher than Pearson dissimilarity (but $p = 0.051$). Notably, the geodesic distance, with segment lengths as short as 300 TRs, outperformed the best accuracy using Pearson dissimilarity which was obtained with the full time course (four times more data; $p < 10^{-4}$; reference $\alpha = 0.05/11 = 0.0045$ given 11 segment lengths).

4.2.5 Identification accuracy and time course length: task data

Although accuracy increased with segment length for *resting state*, length did not predict performance straightforwardly (Figure 4.6A). In particular, *working*

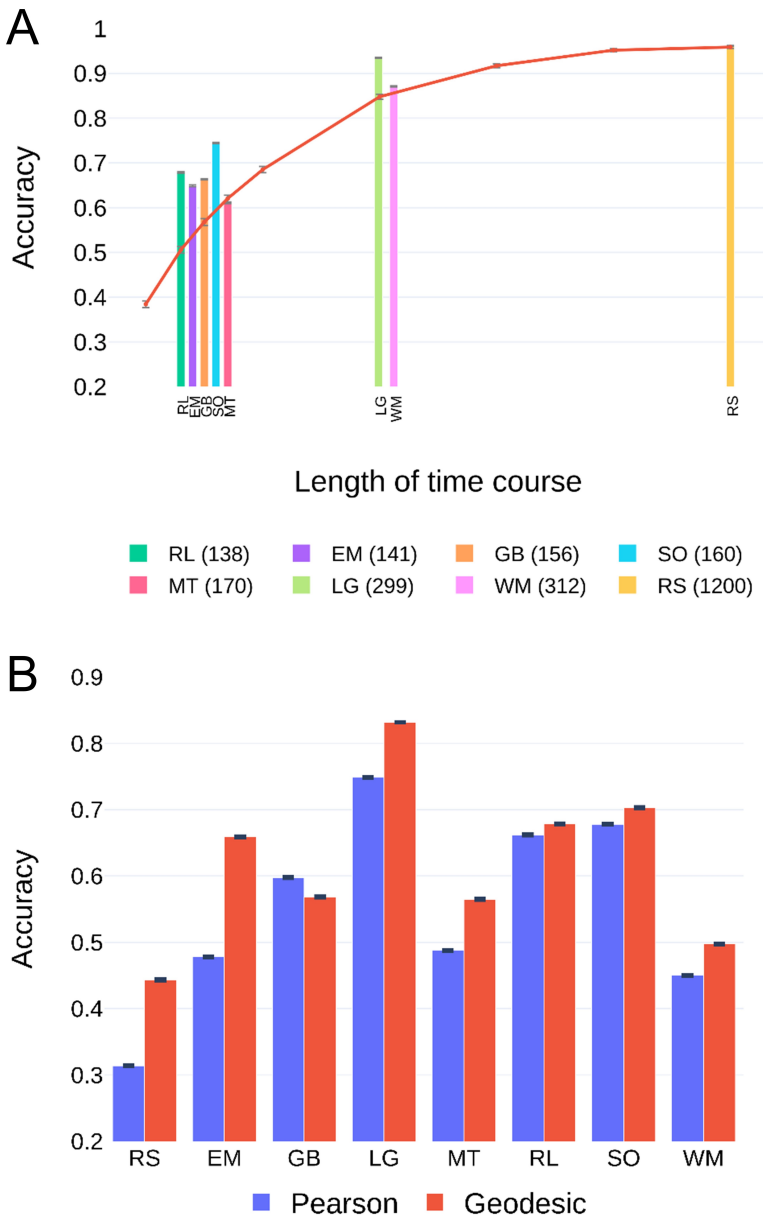

Figure 4.6: Participant identification and time course length. (A) Accuracy based on geodesic distance for *resting-state* and task conditions (time course length in TRs in the inset). The red curve shows the accuracy for *resting-state* data trimmed to segment lengths shorter and longer than those of task data (lengths from left to right: 100, 125, 145, 170, 200, 300, 600, 900, and 1200 TRs). (B) Accuracy when data was trimmed such that all conditions had the same time course length (138 TRs). Error bars indicate standard error of the mean across bootstrap iterations. Abbreviations: EM, emotion processing; GB, gambling; LG, language; MT, motor; RL, relational processing; RS, resting-state; SO, social cognition; WM, working memory.

memory and *language* tasks had comparable time course lengths, but identification accuracy differed by as much as 10%. To probe this issue further, runs were trimmed so that they all had the same duration (138 TRs, which was the length of the shortest task; for conditions with more data, this target length was obtained by deleting time points at the beginning and end of the data segment, thereby retaining the middle part).

With time course length equated, accuracy still varied considerably across tasks (Figure 4.6B). Accuracy obtained using the geodesic distance exceeded that of Pearson dissimilarity for all conditions except the *gambling* task ($p = 1$ for *gambling*, $p < 10^{-4}$ for all other tasks; reference $\alpha = 0.05/8 = 0.00625$ given 8 conditions; Fig C.8). Notably, although *resting-state* had the highest identification accuracy when the entire time course was used, it had the lowest identification accuracy when length was equated across conditions.

4.2.6 Brain subnetworks and participant identification

Particular brain subnetworks are known to be engaged more prominently, as well as exhibit enhanced functional connectivity, during particular tasks [2]. To evaluate performance based on subsets of regions, ROIs were grouped into seven subnetworks (Methods 4.1.3). Was the best subnetwork for identification dependent on condition? Data for all conditions were trimmed so that they had the same length (138 TRs; the `limbic` subnetwork was excluded because identification accuracy was less than 10% across conditions).

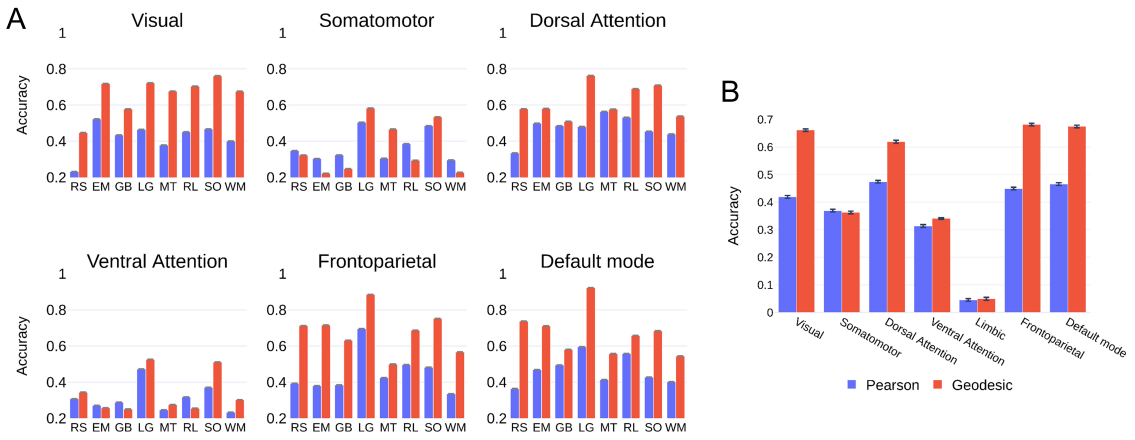

Figure 4.7: (A) Participant identification accuracy using subnetworks. Runs were trimmed such that all conditions had the same time course length. Some subnetworks were more suitable than others for identifying individual differences. The use of geodesic distance showed considerable improvements in accuracy for most subnetworks. (B) Across subnetworks, average participant identification accuracy is displayed. The geodesic distance substantially improved identification accuracy. Error bars indicate standard error of mean across bootstrap iterations. Abbreviations: EM, emotion processing; GB, gambling; LG, language; MT, motor; RL, relational processing; RS, resting-state; SO, social cognition; WM, working memory.

Using geodesic distance improved the accuracy across most conditions for most subnetworks (Figure 4.7A). In particular, for the `visual`, `dorsal attention`, `frontoparietal` and `default mode` subnetworks, accuracy was comparable to that obtained with the whole cortex. For example, the `default mode` subnetwork produced accuracy over 90% for the *language* task. The `frontoparietal` performance on *resting-state* and *emotion processing* was close to 80%. Further inspection of Figure 4.7A revealed additional features of condition/subnetwork combinations. For example, the `visual` subnetwork was not very suitable for identification based on *resting-state* data. Not surprisingly, the `default mode` subnetwork performed well with *resting-state* data. Interestingly, the `frontoparietal` subnetwork performed nearly as well with *resting-state* data, too. These two subnetworks obtained even

higher identification accuracy during the *language* task.

To further evaluate performance of subnetworks, identification accuracy was averaged across conditions (Figure 4.7B). By using the geodesic distance, accuracy improved substantially, with several subnetworks improving by over 20%. Except for the `somatomotor` subnetwork, using the geodesic distance resulted in improved performance ($p = 0.996$ for `somatomotor`, $p < 10^{-5}$ for all other subnetworks; reference $\alpha = 0.05/7 = 0.0071$ given 7 subnetworks; see Figure C.9 for bootstrap distributions). The highest mean accuracies were observed in the `visual, dorsal attention, frontoparietal`, and `default mode` networks for both geodesic and Pearson measures, indicating that some subnetworks are more suitable than others for participant identification.

Figure 4.8 displays geodesic identification accuracy for each condition as a function of subnetwork size. Whereas the smallest subnetwork (`limbic`) performed poorly for all conditions, accuracy did not always increase with size. For example, the `dorsal attention` and `ventral attention` subnetworks have the same size, but the former produced considerably higher accuracy on each condition ($p < 10^{-12}$ for all conditions; reference $\alpha = 0.05/8 = 0.00625$ given the 8 conditions; see Figure C.10 for bootstrap distributions). Across conditions, the `dorsal attention` improved over the same-sized `ventral attention` by over 20%. Of note, the `somatomotor` subnetwork was larger than all but the `default mode` subnetwork, but it produced relatively low identification accuracy; at the same time, the largest subnetwork (`default mode`), was associated with consistently high accuracy across conditions. Finally, no single subnetwork exhibited the highest accuracy for all con-

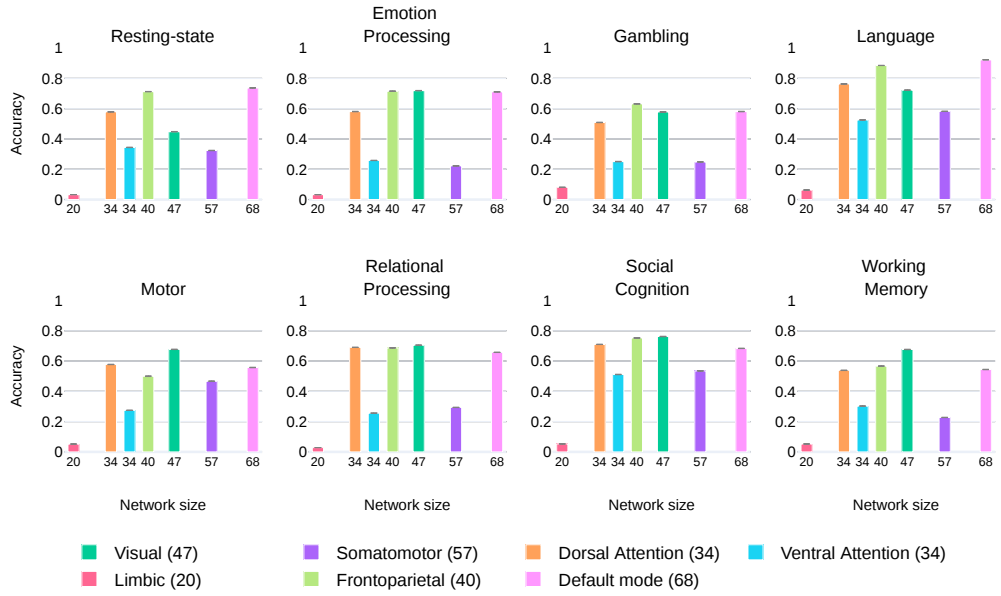

Figure 4.8: Participant identification accuracy plotted against subnetwork size for each condition (geodesic distance). The size of the subnetwork (the number of ROIs) is also indicated in the inset. The error bars represent standard error of the mean across bootstrap iterations.

ditions. In fact, performance varied across conditions, but also varied in particular ways across subnetworks for each condition. Notably, the `visual`, textttdorsal attention, `frontoparietal`, and `default mode` subnetworks performed consistently well. Similar trends were observed for the Pearson dissimilarity measure but overall accuracy levels were lower (Figure C.11).

4.2.7 Combining subnetworks improved identification accuracy

As described, subnetworks had comparable (and sometimes higher) identification accuracy than whole-cortex performance, although subnetworks were associated with much smaller matrices, of course. Could particular subnetworks be com-

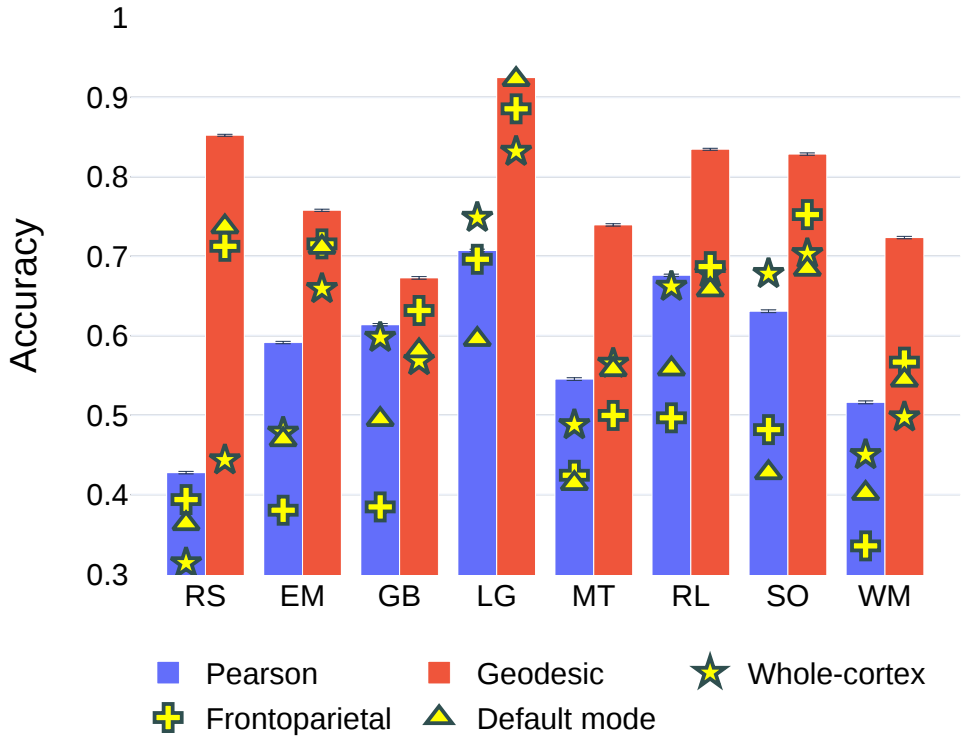

Figure 4.9: Participant identification accuracy by combining subnetworks. For the geodesic distance, the `frontoparietal` (`subnet1`) and `default mode` (`subnet2`) subnetworks were combined. For the Pearson dissimilarity measure, the `dorsal attention` (`subnet1`) and `default mode` (`subnet2`) subnetworks were combined (the top two subnetworks based on mean accuracy across conditions for this measure). Abbreviations: EM, emotion processing; GB, gambling; LG, language; MT, motor; RL, relational processing; RS, resting-state; SO, social cognition; WM, working memory.

bined to further improve identification? We tested this possibility by targeting two subnetworks that exhibited high performance overall, namely `frontoparietal` and `default mode` (see Figure 4.7B). The combined network included all within-network functional connections of course, but also all between-network links (for example, a functional connection between a region of the `frontoparietal` network and a region

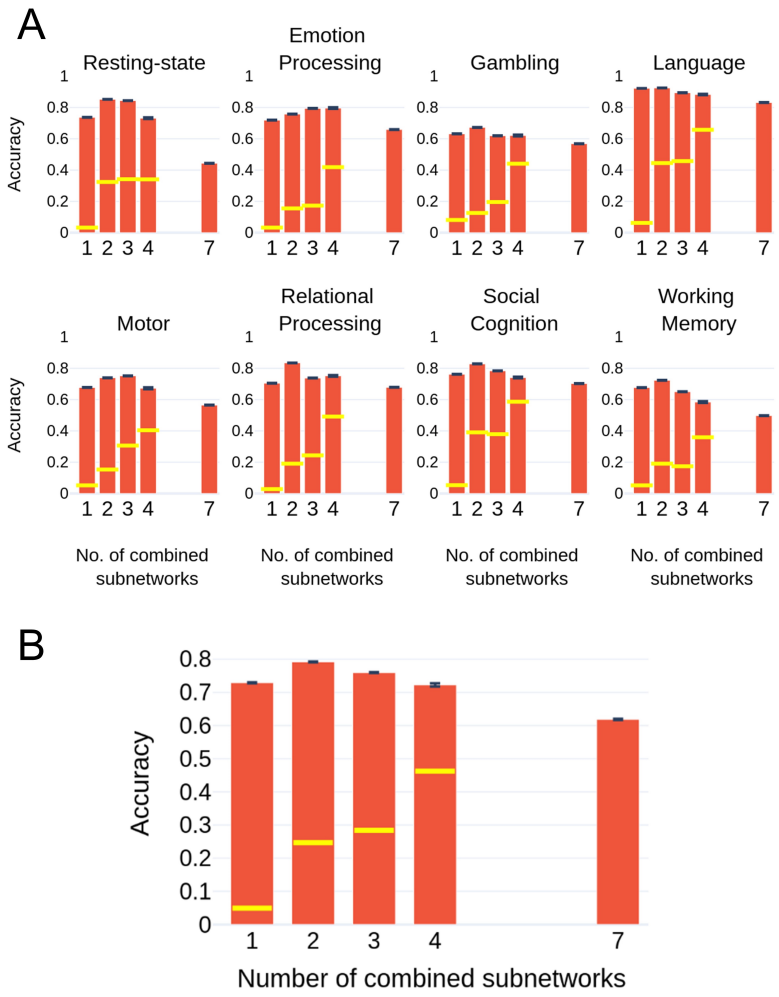

Figure 4.10: Combining up to seven subnetworks. (A) Participant identification accuracy using geodesic distance as a function of the number of subnetworks for each condition. For a particular condition and number of combined subnetworks, the maximum identification accuracy across all combinations of subnetworks is shown with the red bar (the minimum is indicated by the yellow line). Accuracy initially increased with the number of subnetworks but then decreased, and was lowest using whole-cortex FCs (i.e, number of combined subnetworks = 7). (B) Participant identification accuracy averaged across conditions is displayed against number of combined subnetworks.

of the `default mode` network). Time course length was equated for all conditions as in Section 4.2.5. Accuracy using geodesic distance was superior to Pearson dissimilarity (Figure 4.9; $p < 10^{-15}$ for all conditions; reference $\alpha = 0.05/8 = 0.00625$ given 8 conditions; Figure C.12).

Using geodesic distance, the combined subnetwork also outperformed both the individual subnetworks on all conditions except the *language* task ($p = 0.24$ for the *language* task, $p < 10^{-12}$ for all other conditions; $\alpha = 0.05/16 = 0.003125$ given 8 conditions and comparisons with two subnetworks; see Figure C.13-C.14 for bootstrap distributions). In addition, for the geodesic distance, the combined subnetworks exhibited higher accuracy than whole-cortex FC matrices ($p < 10^{-12}$ for all conditions; $\alpha = 0.05/8 = 0.00625$ given 8 conditions; see Figure C.15 for bootstrap distributions) although the number of ROIs in the combined subnetwork (108) was nearly a third as those in the cortex (300). Clearly, the improvement in accuracy was not a simple consequence of increased size, but resulted from improved identity characterization.

To understand whether addition of other subnetworks to the combined network further improved accuracy, we performed identification using combinations of the seven networks taken two, three, four, five, or six at a time. The maximum identification accuracy across all combinations of subnetworks is displayed against the number of combined subnetworks in Figure 4.10A. The minimum identification accuracy across the combinations of subnetworks is also indicated. For all conditions, accuracy initially increased as more subnetworks were considered but then decreased. Performance peaked at 2 or 3 subnetworks for all conditions. Accuracy varied across the combinations of subnetworks (when the number of subnetworks was held constant), and the minimum value (shown in yellow) was less than half the maximum when less than four subnetworks were combined. In Figure 4.10B, identification accuracy was averaged across conditions and displayed as a function

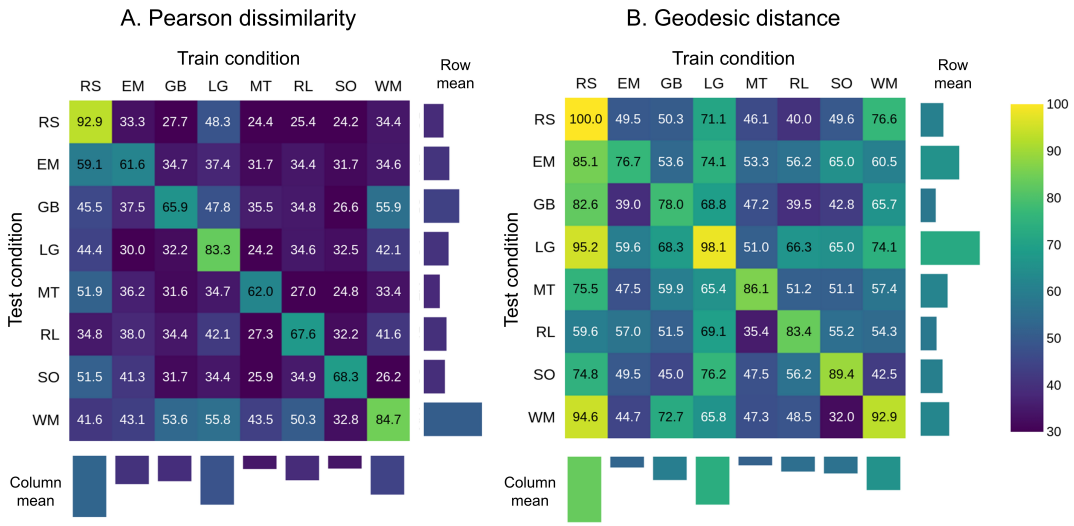

Figure 4.11: Participant identification accuracy when the training and testing data were based on different conditions. The combined network containing the `frontoparietal` and `default mode` subnetworks was employed. The mean accuracy for each train and test condition is also indicated. For example, when *resting-state* is used as training data, the column mean is computed as the accuracy across all other conditions (i.e., except *resting-state* itself). The row means are computed in a similar fashion by excluding the diagonal term. Abbreviations: EM, emotion processing; GB, gambling; LG, language; MT, motor; RL, relational processing; RS, resting-state; SO, social cognition; WM, working memory.

of the number of combined subnetworks.

4.2.8 Transfer of identifiability between conditions

In the previous sections, training and testing data were based on the same condition. Here, we sought to understand if participants could be identified if the training and testing data were obtained from different conditions; for example, identifying a participant performing a *working memory* task when the training used *resting-state* data. Time series length was not equated across conditions because our goal was to evaluate how transferable identity-related information was between pairs of conditions. Accordingly, we did not want to potentially degrade FC informa-

tion by using shorter data segments. Identification was performed on the combined `default-plus-frontoparietal` network, which as discussed performed well across conditions (Figure 4.9).

Results for both geodesic distance and Pearson dissimilary are displayed in Figure 4.11. Whereas Pearson dissimilarity was useful in identifying participants when they performed the same task (within-conditions, diagonal entries), performance deteriorated when the training and test data originated from different tasks. Notably, across-condition identification was considerably higher with the geodesic distance, and this enhancement was rather striking when the training data was from *resting-state*, and to some extent based on the *language* and *working memory* tasks. For example, testing *working memory* data based on training with *resting-state* data yielded 94.6% accuracy, which intriguingly was even better than when training with *working memory* itself (accuracy: 92.9%, $p < 10^{-4}$). On average, training with *resting-state* yielded 83.4% accuracy when testing on *other* conditions (see the "column mean" in Figure 4.11). The present results indicate that the geometry of FC is especially important for across-task identification (see Discussion).

Because in this section time course length was not equated across conditions, we note that those with longer lengths aided across-task identification. Accordingly, transfer might particularly benefit from employing training sets with longer data segments. Nevertheless, future research should also evaluate transfer effects when longer data segments are available for a wider range of tasks (for example, ≥ 300 TRs) so as to characterize their transfer potential.

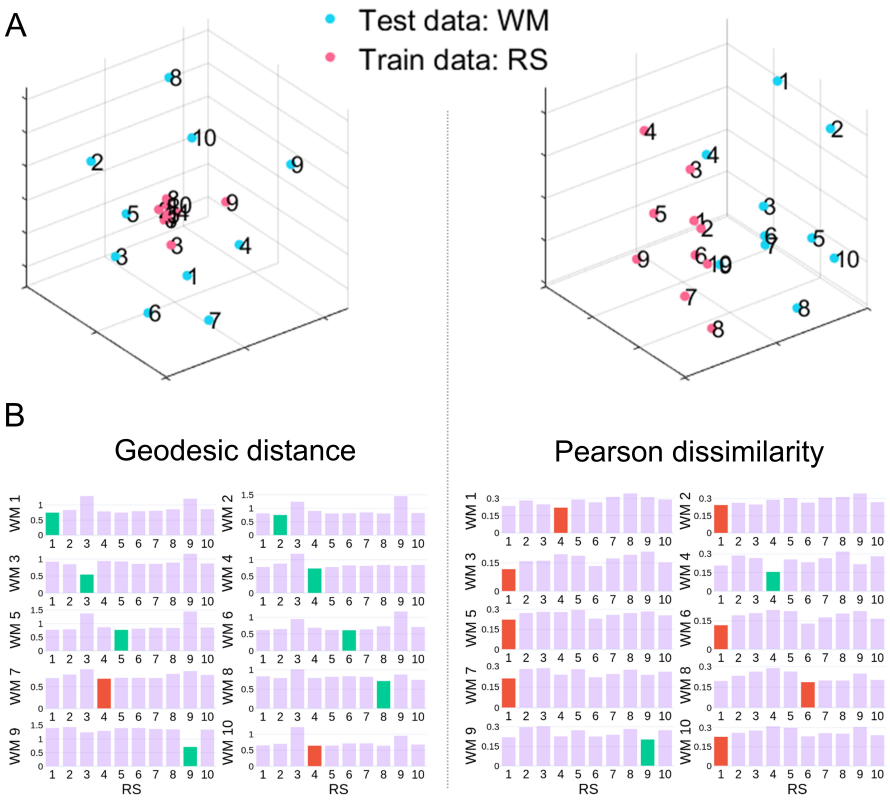

Figure 4.12: Visualization of task and *resting-state* functional connectivity distances/dissimilarities in a three-dimensional space using multidimensioanl scaling. The numbers indicate participant IDs. (A) Distances/dissimilarities between the functional connectivity matrices of *resting-state* (RS, used for training) and *working memory* (WM, used for testing) for a set of 10 randomly chosen participants. Online figures are available [129]. (B) Participant-level distances/similarities between training and testing data. Correct identification is marked in green and incorrect in red. For example, when using geodesic distance, the best candidate for WM participant 1 (call it WM1) was RS participant 1 (RS1), and the best candidate for WM2 was RS2. However, incorrect classifications were also observed, such as RS4 (not RS7) being closest to WM7. For Pearson dissimilarity most classifications were incorrect, such as RS1 (not RS10) being most similar to WM10. Distances based on the two measures have arbitrary units, and are not comparable across them.

4.2.9 FC geometry of task and resting-state data

As some conditions yielded high identification accuracy when training and testing were based on different conditions, we sought to visualize distance/dissimilarity in a lower dimensional space via multidimensional scaling. Figure 4.12A displays the

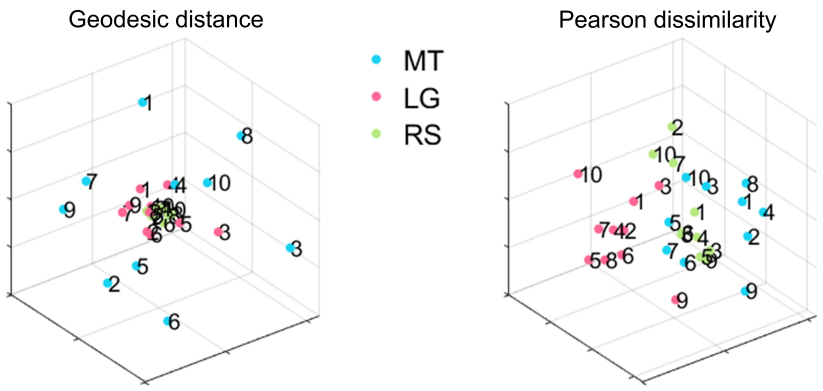

Figure 4.13: Functional connectivity geometry of *resting-state* and task conditions (online figures are available [129]). Training data for 10 random participants were employed (indicated by the numbers). Distances/dissimilarities in low dimensions were obtained via multidimensional scaling. Note that the geometry in low dimensions differed considerably for geodesic and Pearson, suggesting that condition categorization (not participant identification) should capitalize on such geometry for better performance. Abbreviations: MT, motor; LG, language; RS, resting-state.

low-dimensional representation of the distances/dissimilarities for a set of randomly chosen participants when *resting-state* was employed for training data and working memory for testing (untrimmed data). Based on the geodesic distance, *resting-state* FC matrices were relatively close together to one another; in contrast, *working memory* FC matrices were further "spread out". Intriguingly, such geometry allowed for the separation of FCs based on participant identity. To see this, consider the panels in Figure 4.12B, which show participant-level distances. In contrast, using Pearson dissimilarity, the geometry did not allow accurate participant identity. In fact, nearly all participants in this illustrative sample were misidentified.

The results in Figure 4.12A prompted us to investigate, in an exploratory fashion, distance/dissimilarity between conditions, specifically, *resting-state*, *motor*, and *language* (Figure 4.13). Intriguingly, the geometry of distances was quite different when geodesic distances were used compared to Pearson dissimilarity. These

observations suggest that when FC matrices are used for *task classification* (not identification as done here), different algorithms may be more suited for this aim. For example, non-linear radial basis functions might function better for the geodesic case, and linear classifiers for Pearson dissimilarity. Although a fuller investigation of this issue is beyond the scope of the present paper, we believe this is a fruitful direction. Furthermore, the analysis of functional connectivity of mental states should take into account participant-related information since it plays a potentially dominant contribution in the identification of mental states [132].

4.3 Discussion

In this paper, we investigated participant identification based on FC matrices from fMRI data by employing geometry-aware methods. Correlation matrices are objects that lie on non-linear surfaces, and thereby benefit from non-Euclidean distance measures. Indeed, we show that using the geodesic distance improved participant identification, at times by as much as 20%. Further, low-dimensional visualization based on geodesic distance contributes to understanding how FC geometry affects identification.

4.3.1 Factors influencing participant identification

Scan duration determines the amount of data used to estimate FC matrices, and played a key role in identification accuracy (see Figure 4.5). For *resting-state* data, accuracy improved with time course length and was close to 95% when the

entire data were employed (1200 TRs), but fell to under 50% when trimmed to under 150 TRs. The steep drop is possibly due to the underlying dynamics of *resting-state* data [40], and reveals that longer data segments are required to more robustly identify functional connectivity patterns that are unique to individuals. Notably, inspection of Figure 4.5 indicates that accuracy using Pearson dissimilarity increased very modestly despite substantial increases in data length. If such trends can be extrapolated, it would suggest that it is unlikely that accuracy with Pearson dissimilarity would reach that obtained using geodesic distance. Conversely, using the geodesic distance resulted in higher accuracy than Pearson dissimilarity even when, say, only a fourth of the data were employed for FC estimation. Thus, a more suitable geometry is particularly appealing when data-limited scenarios are envisioned.

When time course length was trimmed to the same duration, identification accuracy still varied across scanning conditions. The *resting-state* condition resulted in the lowest accuracy. With the data trimmed to the minimum amount of data, the *language* task exhibited over 80% accuracy. Accuracy of all task conditions exceeded 50%, with four of them exceeding 60%. Thus, even with rather limited amounts of data identifying the participant was considerably better than the chance level of 1%. In addition, we observed considerable variability is performance across conditions, consistent with previous literature suggesting that brain states can be manipulated to emphasize individual differences in FC [108].

Thus far, we have discussed findings based on whole-cortex FC matrices (300 ROIs were employed). We reasoned that particular subsets of regions potentially

might be more informative than others. To evaluate this possibility identification was applied to *resting-state* and task conditions separately for each individual subnetwork of the Yeo parcellation [94]. The FC matrices employed were therefore relatively small (the number of ROIs ranged from 20-68). Four subnetworks (`vision, dorsal-attention, frontoparietal, default`) stood out as consistently exhibiting the highest levels of performance. The average accuracy across conditions approach 70% for the four networks. Intriguingly, accuracy for the *language* task based on the `frontoparietal` and `default` subnetworks exceeded that observed with the whole cortex. Whereas subnetwork size might contribute to its ability to identify participants, it is clearly not the driving factor. For example, the `dorsal-attention` and the `ventral-attention` networks had the same number of ROIs, but the former outperformed the latter consistently (on average by over 30%).

To further explore subnetwork contributions we also combined the two that displayed the highest individual accuracy (`frontoparietal` and `default`) into a single network. Remarkably, the combined network always numerically outperformed the individual subnetworks, and indeed the entire cortex. When additional subnetworks were combined, accuracy initially increased but then decreased. Accuracy peaked at two or three subnetworks, with whole-cortex FCs having the worst performance across conditions. Accuracy also varied across combinations of subnetworks, with the minimum value less than half that of the maximum when less than four subnetworks were combined. These results are related to the non-uniformity of within-subject test-retest reliability of connectivity profiles, and might inform how individual differences are associated with heritability and cognitive ability [133].

Thus, future work on individual differences using connectomes should not only consider tasks but also choose appropriate measures and subnetworks that emphasize these differences.

Although it was beyond the scope of the present study, it would be valuable to investigate in future studies factors contributing to the performance of individual subnetworks, and their combinations. For example, subnetworks may contribute highly to identification because their individual-specific functional connectivity information capitalizes on the contributions of these subnetworks to task performance. Alternatively, but not mutually exclusively, subnetworks that do not participate as much during a task may contain diagnostic information with respect to participant identity.

To what extent does participant identification transfer between experimental conditions? We found that training with one condition and testing with another produced good levels of identification accuracy. Certain combinations that on the surface were not obvious produced particularly impressive results; for example, training with *gambling* and testing with *working-memory*, or training with *working-memory* and testing with *language*. Training with *motor* produced the least transfer to other tasks, perhaps due to the low-level specificity of this task. Notably, training with *resting-state* produced very high transfer, such that testing with each task attained accuracy over 75% (with the exception of *relational processing*), and in some instances over 90%. The choice of measure was particularly important for transfer of identifiability and accuracy, with *working-memory* attaining nearly 95% using geodesic distance but less than 42% using Pearson dissimilarity.

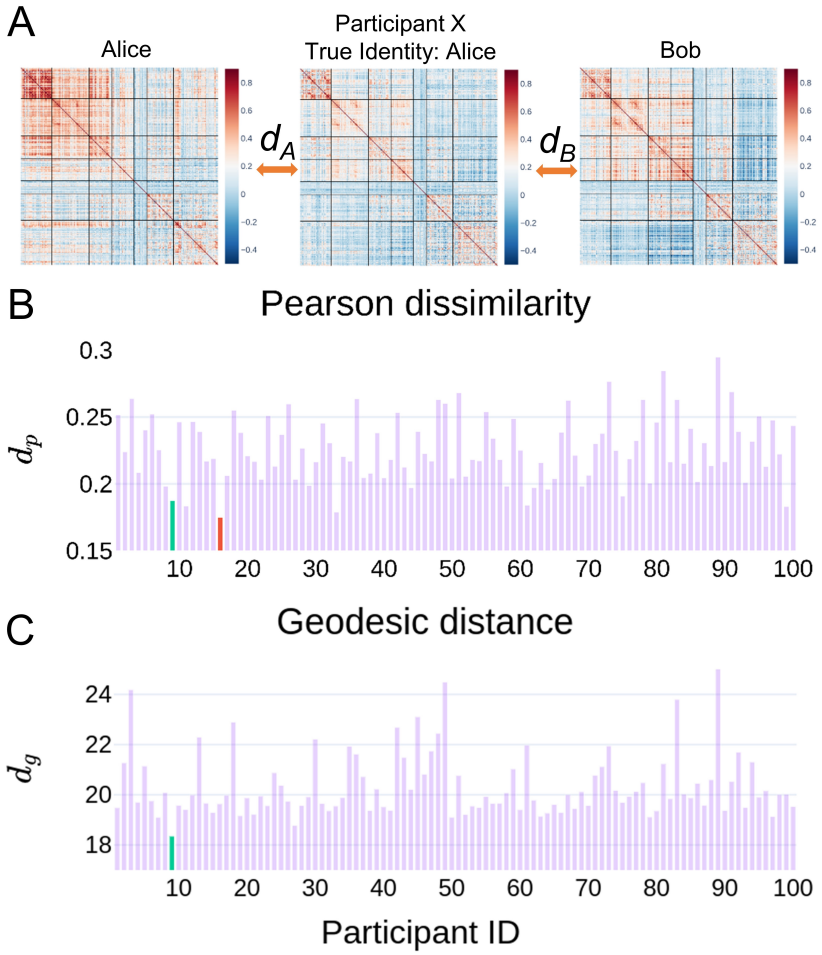

Figure 4.14: Visual comparison of functional connectivity (FC) matrices can be unintuitive. (A) Example FCs from *resting-state* data where the geodesic distance correctly labeled the test participant but Pearson dissimilarity did not. Pearson dissimilarities and geodesic distances between the test-FC and each of the FCs in the training data are shown in (B) and (C). The green bar indicates the distance between the test-FC to the correct training set FC; the red bar indicates an incorrectly labeled training set FC. For the geodesic distance, the labeled participant had indeed the smallest value; not so in the case of Pearson dissimilarity. This example also questions the common practice of informally evaluating functional connectivity similarity via simple visual inspection. At the very least, it is not immediate that participant X is more similar to Alice than Bob.

4.3.2 Low-dimensional distance visualizations

Relationships between high-dimensional FC matrices (300×300) were visualized in three Euclidean-space dimensions using multidimensional scaling. Both

the Pearson dissimilarity measure and geodesic distance were used. Note that computing geodesic distances takes into account the non-linear geometry of correlation matrices. Once their distances are computed, and the space nonlinearity taken into account, they can be illustrated in Euclidean space (naturally, some distortion ensues due to dimensionality reduction).

In our explorations, low-dimensional visualizations reflected identification accuracy on the full data, and thus preserved important distance information. In particular, the higher identification accuracy using the geodesic distance resulted in relatively low within- and high between-participant distances. Visualization of FC from task data revealed insights into the geometry of task correlation matrices in relation to *resting-state*. Identification accuracy is related to the ratio of within- to between-participant distances. Surprisingly, with geodesic distances, tasks associated with higher identification accuracy exhibited smaller between-participant distances. Still, the more favorable ratio of within- to between-participant distances led to favorable identification accuracy. Thus, the underlying geometry of functional connectivity may provide further insights into our finding that high identification accuracy was attained when training and testing were based on different scanning conditions.

In the visualizations based on geodesic distance, distances between task FCs did not appear to form convex sets (if A and B are two points in a convex set, any point on the line joining them also belongs to the set), and were instead in clustered arrangements. Of note, previous work performing clustering of FCs [40, 134] have used k-means which are not well suited to finding non-convex clusters [135]. Instead,

methods such as spectral clustering [136] and non-linear support-vector kernels [137] are capable of capturing very general structures, and are potentially more suitable for classifying functional connectivity.

Pearson correlation is a common approach to compare FC matrices. The present study demonstrates that non-linear measures are better suited to characterize functional connectivity relationships. The low-dimensional visualization briefly explored here hints at the different geometries associated with the geodesic non-linear metric and the Pearson approach. Surprisingly, we noted in our investigations that simple visual inspection of the correlation matrices as commonly done in the field to highlight similarities between conditions can also be problematic, and in fact can lead to unintuitive scenarios (Fig 4.14).

Chapter 5: Concluding Remarks

In this dissertation, we characterized dynamic spatiotemporal patterns and connectivity in resting-state, task, and naturalistic fMRI data.

In Chapter 2, we developed an approach employing reservoir computing, a type of recurrent neural network, and show the feasibility and potential of using it for the analysis of temporal properties of brain data. The framework was applied to both Human Connectome Data and data acquired while participants viewed naturalistic movie segments. We show that reservoirs can be used effectively for temporal fMRI data, both for classification and for characterizing lower-dimensional "trajectories" of temporal data. Importantly, robust classification was performed across participants (in contrast to within-participant classification). We hypothesize that low-dimensional trajectories may provide "signatures" that can be associated with tasks and/or mental states. Code is available at `http://github.com/makto-toruk/brain-esn`.

In Chapter 3, we employed an LSTM-based architecture to characterize distributed spatiotemporal patterns of dynamic movie-watching fMRI data. The latent space of reservoirs contained rich characterizations of the spatiotemporal data utilizing a large reservoir matrix that satisfies the "echo state property". LSTMs are

known to achieve similar or superior performance with fewer recurrent units, but unlike reservoirs, the recurrent weights are optimized during training. The proposed LSTM framework was applied to movie-watching data from the Human Connectome Project where 176 participants watched 15 unique clips. The availability of this larger dataset enabled efficient training of LSTM networks. Representations associated with clips required capturing long-term dependencies, were consistent across participants, and generalized to previously unseen participants.

We obtained low-dimensional representations using a unified framework that simultaneously learned the best latent space for clip classification as well as the low-dimensional space to project on to. Low-dimensional trajectories obtained using LSTMs captured important temporal properties in few dimensions and served as "signatures" for movie clips. Saliency maps uncovered brain regions and their time-varying importance to prediction. Recent work has utilized relatively static characterizations of fMRI data (e.g., functional connectivity) to predict individual differences in behavior and personality. Based on our hypothesis in Chapter 2 that trajectories may be associated with mental states, we used LSTMs to relate spatiotemporal patterns to the behavior and personality of individuals. Across a range of behavioral and personality measures, LSTMs outperformed the state-of-the-art. The results hint that neural responses are composed of stimuli- and individual-related components. Understanding their interaction and intrinsic dimensionality is a fruitful avenue for future research. Code is available at `https://github.com/makto-toruk/lstm-fmri-dynamics`.

The results in Chapters 2 and 3 provide evidence that brain dynamics must

be embraced for a fuller characterization of underlying processes. More specifically, our work shows that to relate neural responses to the stimuli that evoke them, it is essential to characterize distributed spatiotemporal patterns in these responses. We show promising results that suggest the existence of complex but consistent spatiotemporal structure in brain data obtained with fMRI. Application of these techniques to fMRI and other types of brain data obtained during naturalistic and dynamic experimental paradigms can advance our understanding of how the brain integrates diverse processes to support mental functions.

In recent years, time-series correlation matrices or functional connectivity matrices have been widely used to characterize spatiotemporal patterns in fMRI data. For example, they have been used to cluster brain states [40], characterize dynamic functional states [106], perform participant identification [41], and understand how tasks reconfigure brain networks [107]. However, in most of these studies, the geometry of functional connectivity matrices has not been appropriately handled. In Chapter 4, we propose the use of a geodesic distance metric that reflects the underlying non-Euclidean geometry of functional connectivity matrices. We compared identification performance (also called "fingerprinting"; that is, assigning a participant label to novel functional connectivity data) obtained with standard Pearson correlation and the proposed geodesic distance. The latter not only improved identification accuracy but also provided insights into the geometry of task and resting-state conditions. Importantly, the approach advocated we employed is general and can be utilized to study the clustering of brain states, how tasks potentially reconfigure brain networks, and to characterize intersubject correlations. Code is available

at `https://github.com/makto-toruk/FC_geodesic`.

The contents of Chapter 2 and Chapter 4 are published in *Neuroimage* [27,42]. At the time of writing this dissertation, the contents of Chapter 3 were under review.

Chapter A: Supplemental Material for Chapter 2

Type	Run	Name (Year)	Length
Funny	1	Step Brothers (2008)	2:09
	2	Wedding Crashers (2005)	2:03
	3	Night at the Museum: 2 (2009)	3:01
	4	Patch Adams (1998)	1:40
	5	The Elevator (2010)	3:02
	6	Bruno (2009)	2:21
Scary	1	The Eye (2002)	3:01
	2	Paranormal Activity 3 (2011)	2:05
	3	Entry 18 (2009)	2:24
	4	The Blair Witch Project (1999)	3:02
	5	The Others (2001)	2:03
	6	Shutter Island (2010)	1:42

Table A.1: Film names and clip duration for the "scary" and "funny" conditions.

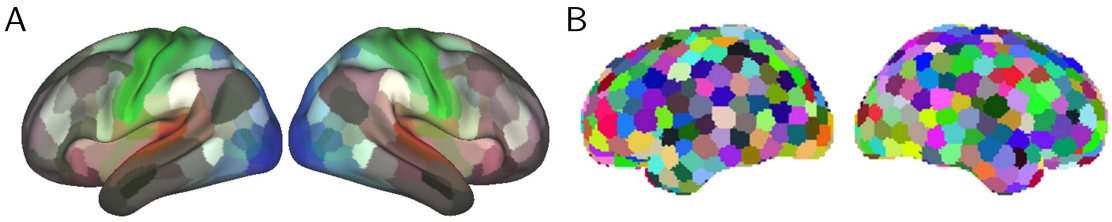

Figure A.1: Region of interest (ROI) masks. (A) For Human Connectome Project data, 360 cortical ROIs as provided by [72] were used. (B) For movie data, 500 cortical ROIs obtained from using k-means clustering on the spatial coordinates $\{x, y, z\}$ of cortical voxels were used. Two amygdala regions (one per hemisphere) were also included but are not show here.

Working memory: "2-back" vs. "0-back"		
	Accuracy on "first" set	Accuracy on "second" set
Reservoirs	86.5%	86.3%
Raw activation	78.9%	77.6%
Concatenation	82.0%	82.8%
Auto-regressive model	81.2%	81.1%
Theory of mind: "social" vs. "random"		
	Accuracy on "first" set	Accuracy on "second" set
Reservoirs	91.8%	91.9%
Raw activation	83.4%	84.2%
Concatenation	85.9%	87.8%
Auto-regressive model	87.2%	86.6%

Table A.2: Comparison of mean cross-validation accuracy using the "first" dataset and classification accuracy on the "second" dataset. The similar results on the two datasets indicate the robustness of the cross-validation scheme.

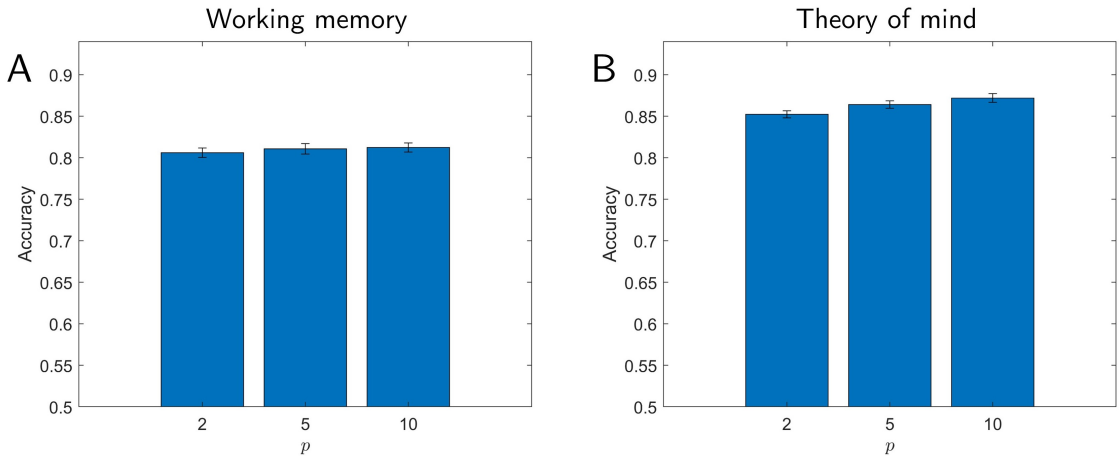

Figure A.2: Classification accuracy using autoregressive models for working memory (A) and theory of mind (B). Results are shown as a function of model order, p.

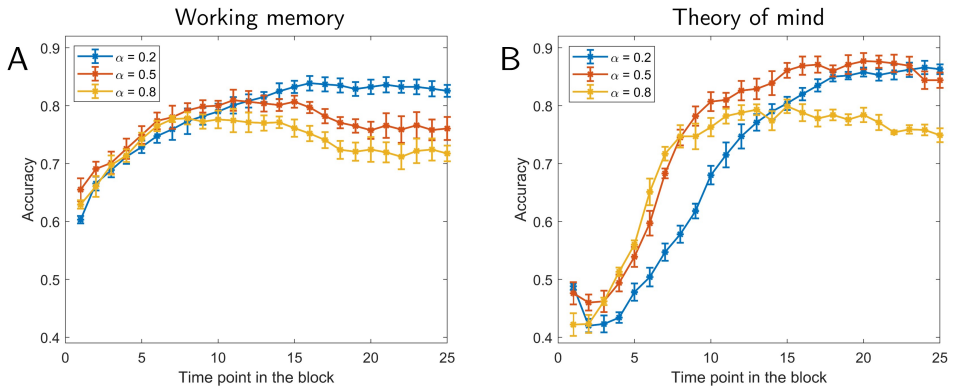

Figure A.3: Classification accuracy as a function of time using only the low-dimensional data (10 top/bottom principal components for working memory, and 12 top/bottom principal components for theory of mind). Results for working memory (A) and theory of mind (B). Accuracy is shown as a function of time point within a task block. Different curves show results for different forgetting rates, α. The values of τ were based on the parameters exhibiting highest accuracy in Fig. 2.3. Error bars show the standard error of the mean.

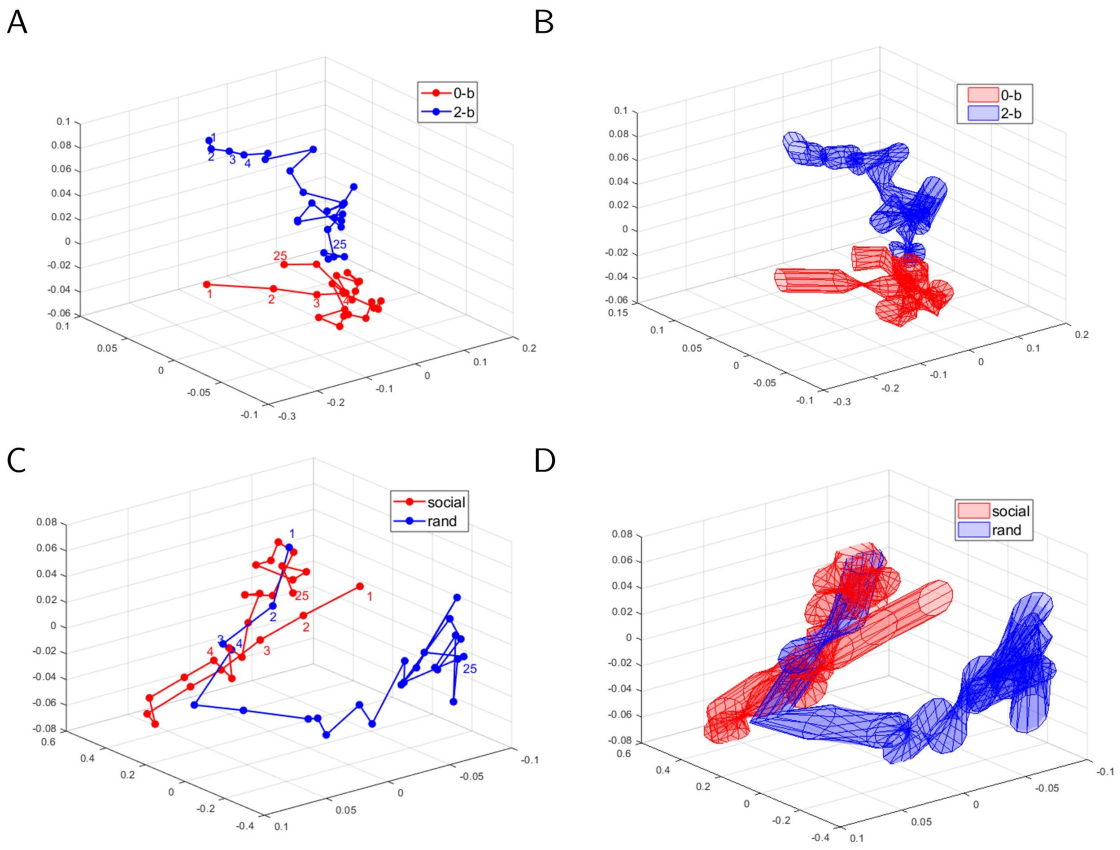

Figure A.4: Temporal trajectories based on the "top" three principal components of the input time series (that is, no reservoir) for fMRI task data. Mean trajectories are displayed in (A) for working memory and (C) for theory of mind. Variability (standard error across participants) is displayed in (B) and (D), respectively. For working memory data (A-B), these trajectories were well separated throughout the block. However, for theory of mind data (C-D), the trajectory for the social condition did not evolve temporally as seen when using reservoirs; note that the final states (see point 25) were close to the initial ones (see points 1-2). Therefore, it appears that a low-dimensional representation based directly on input activations does not adequately capture the temporal evolution structure associated with the social condition.

Chapter B: Supplemental Material for Chapter 3

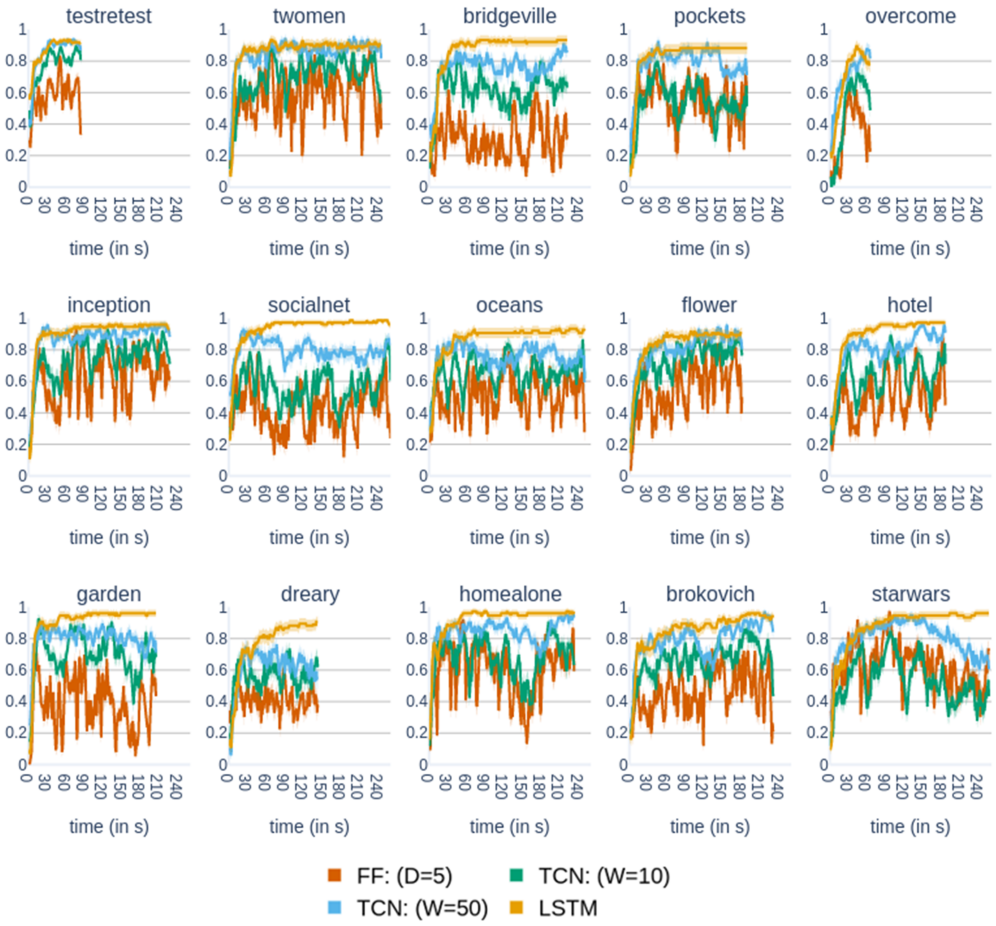

Figure B.1: LSTMs and competing models: feed-forward (FF, 5 layers) classifiers, temporal convolutional networks (TCN, kernel widths of 10 and 50). Since our framework predicted labels at each time step, true positive rate for each clip was determined as a function of time. Error bars show the standard error of the mean across test participants.

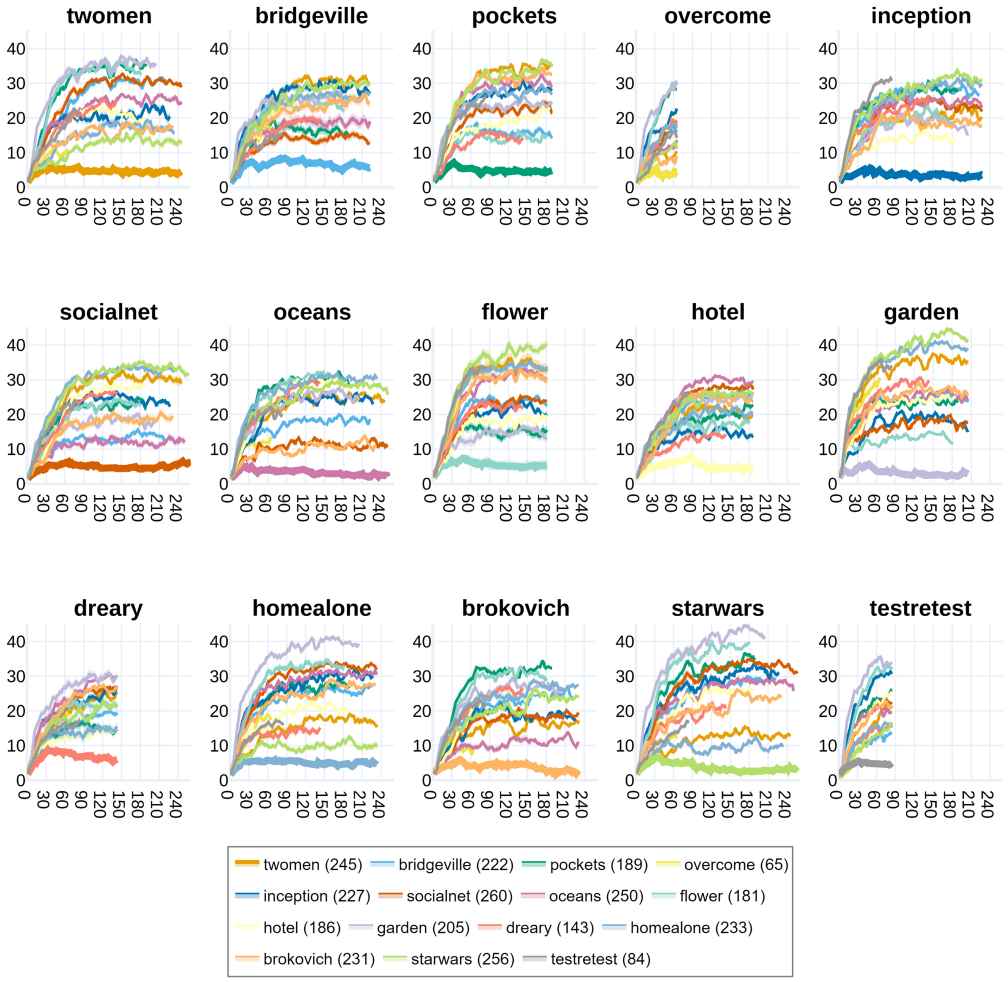

Figure B.2: Distance between trajectories. The distance between the clip trajectory while watching a particular clip (title of each plot) and the mean trajectory across participants for a second clip was computed. Thicker line in each subplot corresponds to the distance of participants' movie trajectories (indicated in the title) to the mean of this clip.

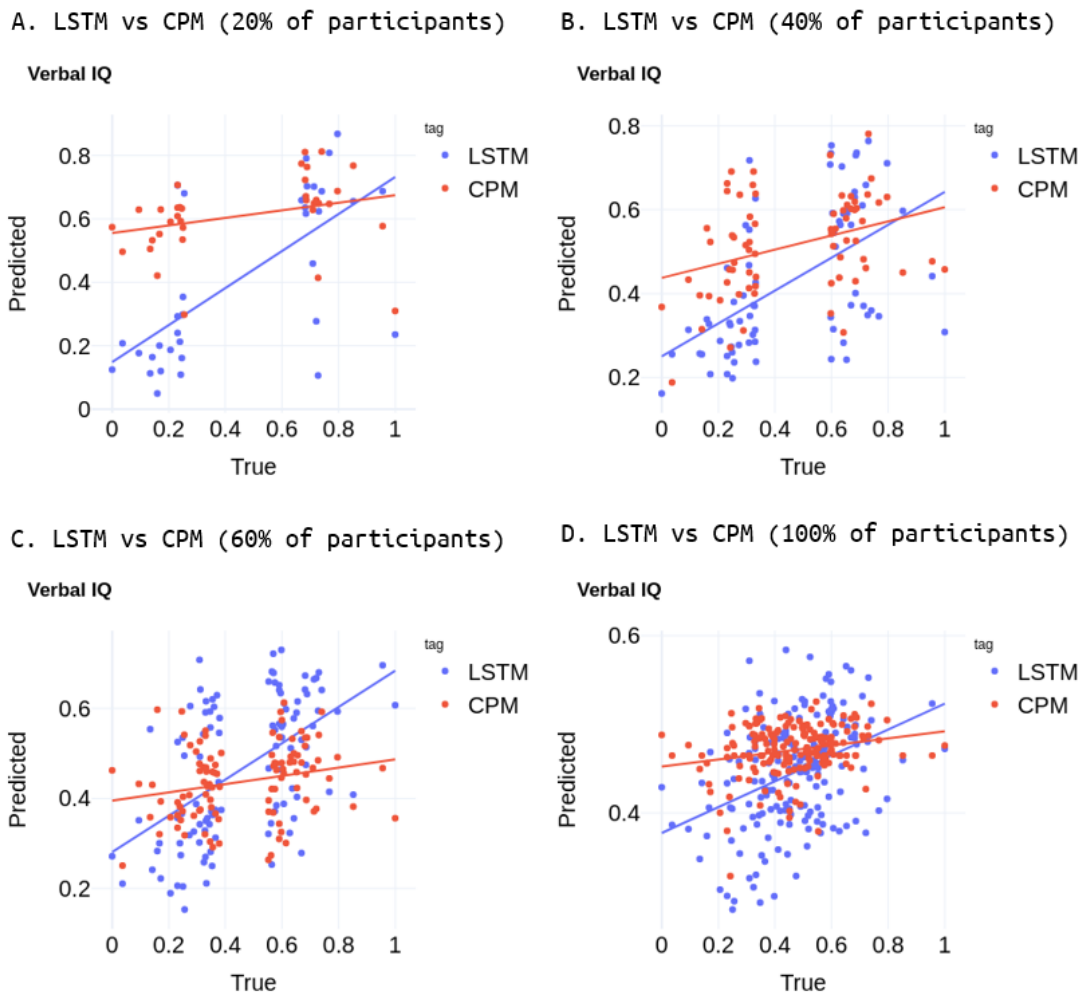

Figure B.3: Correlation between predicted and true verbal IQ scores (rescaled between 0 and 1) while watching *Bridgeville* using (A) top/bottom 20% of the scorers, (B) top/bottom 40%, (C) top/bottom 60%, (D) top/bottom 100%. For all participant cutoffs, predictions based on LSTM outperformed connectome-based predictive modeling (CPM).

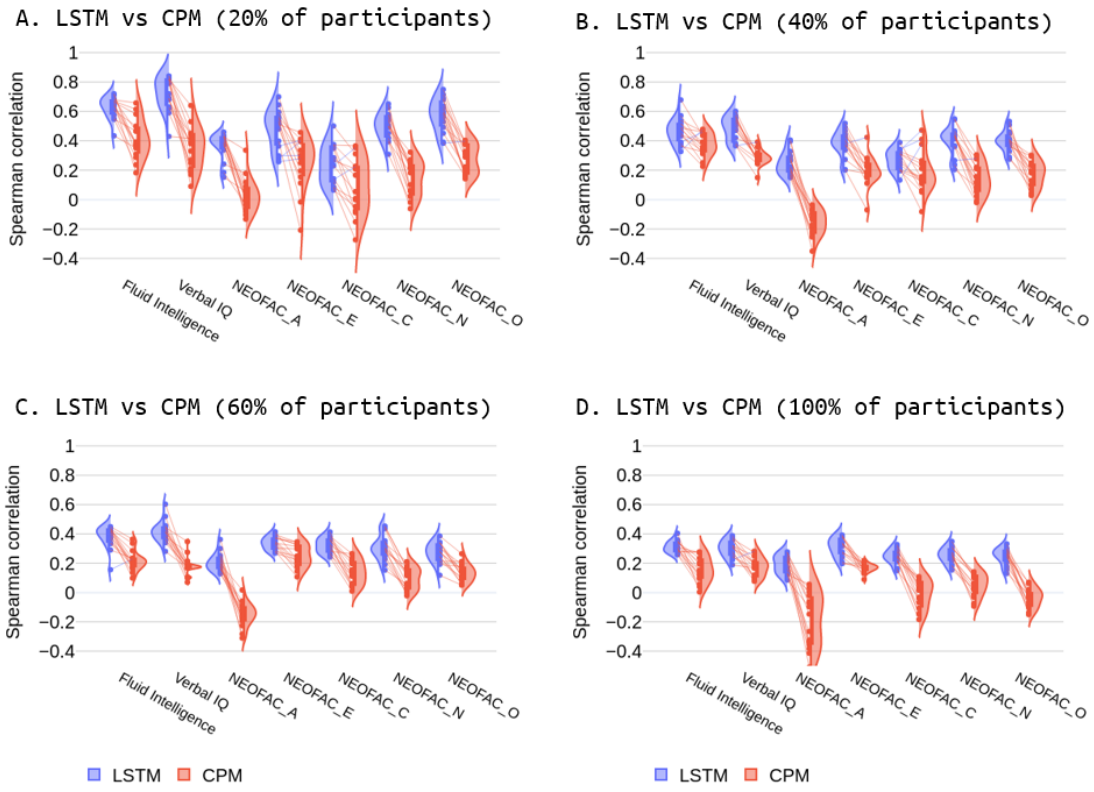

Figure B.4: Comparison of LSTM prediction accuracy with connectome-based predictive modeling (CPM) using (A) top/bottom 20% of the scorers, (B) top/bottom 40%, (C) top/bottom 60%, (D) top/bottom 100%. Connecting lines indicate accuracy for the same clip.

Chapter C: Supplemental Material for Chapter 4

C.1 Identification accuracy when runs were not trimmed

In the main body of the text, to ensure that only task-related segments of a run were retained, "mini resting periods" in the form of fixation periods were removed (see Section 4.1.1). We repeated our analysis without trimming runs of task data using whole-cortex FCs. Identification accuracy for each condition is shown in Fig. C.1. Accuracy obtained using the geodesic distance exceeded that of Pearson dissimilarity for all conditions except the *gambling* and *relational* conditions ($p = 1$ for *gambling* and *relational* conditions, $p < 10^{-6}$ for all other conditions; reference $\alpha = 0.05/8 = 0.00625$ given 8 conditions; Fig. C.2). The mean improvement using geodesic distance was around 8% (as high as 18% on *resting-state* data). Note that even with fixation periods trimmed (original analysis), accuracy with geodesic distance was not clearly favorable for the gambling and relational conditions (see Figs. 4.3 and 4.6B). In general, the precise geometry of task FCs on the positive semidefinite cone is a potential reason why geodesic distance improves participant identification substantially on some tasks but not on others.

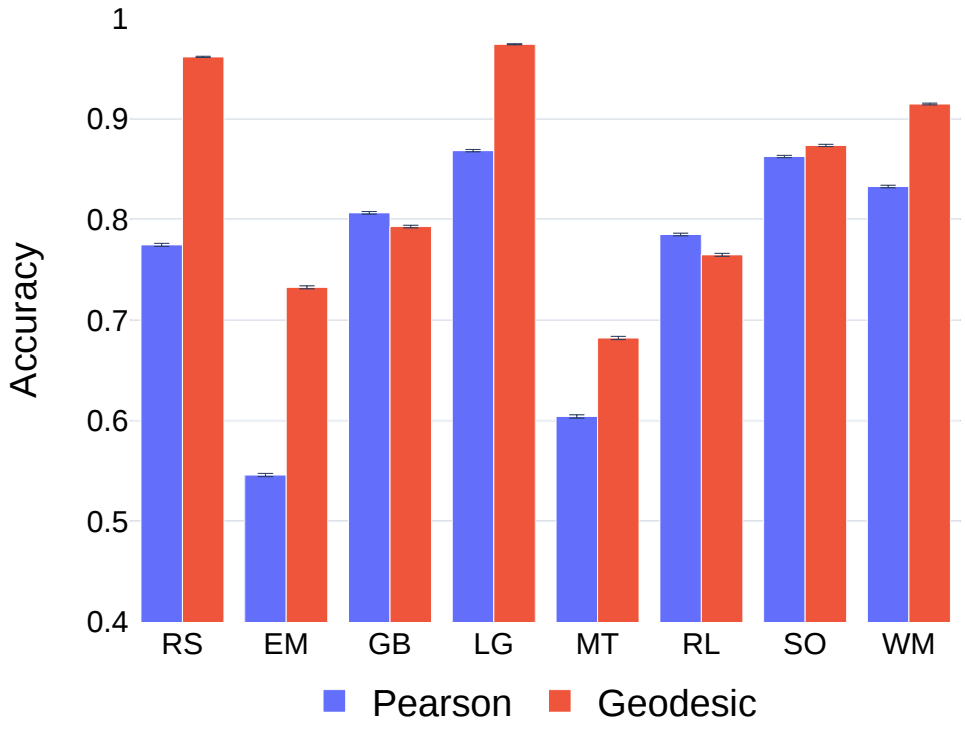

Figure C.1: Identification for the eight conditions using the geodesic distance and Pearson dissimilarity. Training and testing data were from the same condition. Fixation period or "mini resting periods" were not trimmed from runs of task data (as done in Fig. 4.3). Abbreviations: EM, emotion processing; GB, gambling; LG, language; MT, motor; RL, relational processing; RS, resting-state; SO, social cognition; WM, working memory.

C.2 Effect of global signal regression on identification

We repeated our analysis by including global signal regression (GSR) in the preprocessing pipeline for resting-state data [118, 119]. The use of GSR is still debated [138] and can potentially spread underlying group differences to regions that may never have had any [139]. We limit our analysis in this section to *resting-*

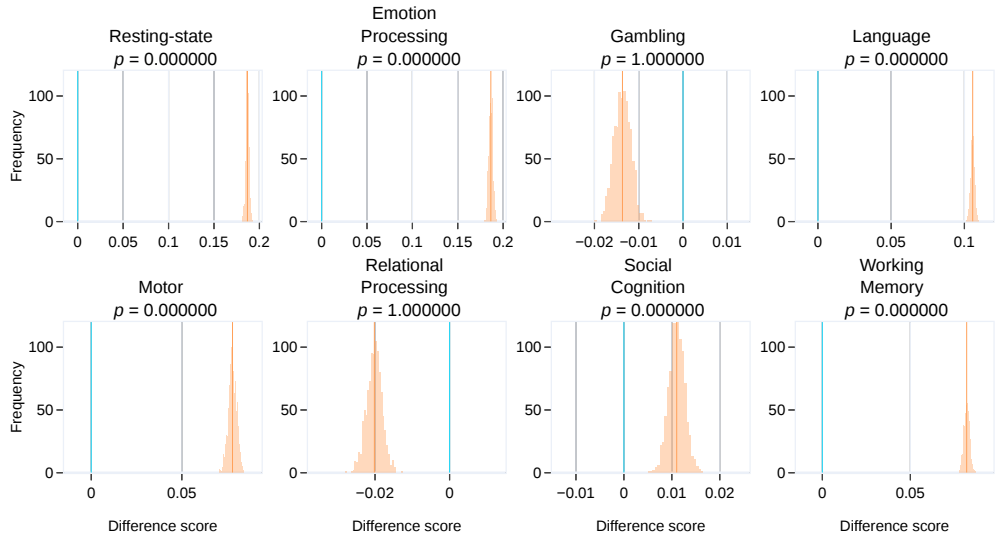

Figure C.2: *Whole-cortex FCs without trimming runs:* Comparison of identification accuracy based on geodesic distance and Pearson dissimilarity for each condition. Identification was based on whole-cortex FCs. Runs were not trimmed as in the main body of the work (see Section 4.1.1). For each condition, the distributions shown in orange represent the difference between the mean participant identification accuracy using the geodesic distance and Pearson dissimilarity across the outer bootstrap iterations (see Section 4.1.7). The orange line indicates the mean of the difference distribution and the blue line indicates zero difference.

state data; we did not include GSR in the preprocessing pipeline for results in the main text. In the data employed (see Acknowledgements), 8 subjects' data were removed because they did not not pass quality control check [119]. Thus, the results reported this section were based on $N = 92$ participants. We performed participant identification using whole-cortex FCs. Regardless of the inclusion of GSR in preprocessing, identification accuracy improved using geodesic distance compared to Pearson dissimilarity (Fig. C.3A). However, using GSR improved accuracy for both measures. When segments of smaller lengths were extracted from *resting-state* data, accuracy improved using geodesic distance for all segment lengths (Fig. C.3B).

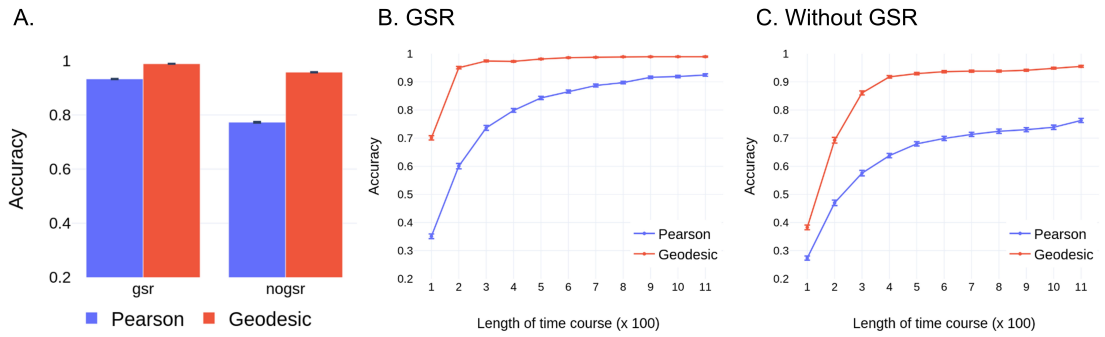

Figure C.3: A. Participant identification accuracy with global mean regression (GSR) included in the preprocessing pipeline (gsr) or not (nogsr). Accuracy using geodesic distance exceeded Pearson dissimilarity for both preprocessing methods. Participant identification accuracy as a function of segment length for *resting-state* data with GSR (in B) and without GSR (in C). In both cases, accuracy using geodesic distance exceeded Pearson dissimilarity at each segment length. Error bars indicate standard error of the mean across bootstrap iterations.

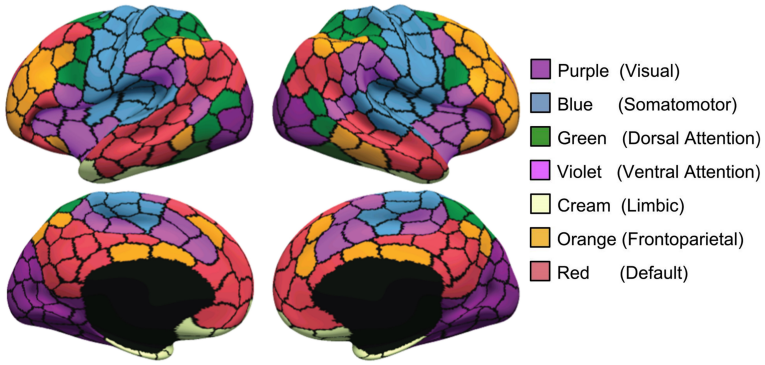

Figure C.4: Parcellation of the cortex into 300 ROIs as provided by [120]. ROIs were grouped into the 7 networks described in [94].

When GSR was used, accuracy using geodesic distance was close to 95% with only 200 time points (compared to 70% without GSR; Fig. C.3C).

C.3 Effect of number of ROIs in the parcellation on identification

To study the effect of the parcellation scheme on participant identification accuracy using the two measures, we employed various parcellations with ROIs ranging

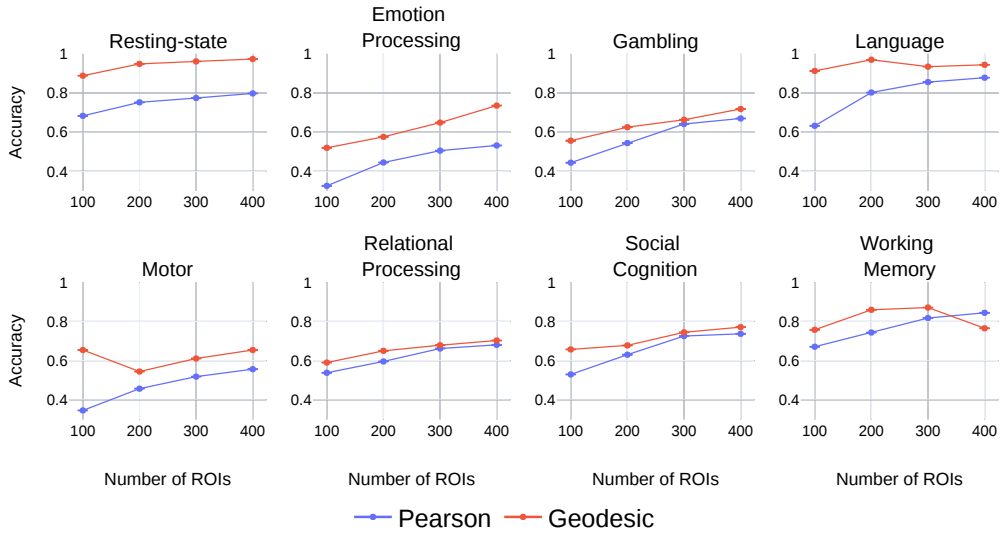

Figure C.5: Participant identification accuracy as a function of the number of ROIs. Here, training and testing data are from the same condition. Error bars indicate standard error of the mean across the bootstrap iterations.

from a 100 to 400. In general, mean participant identification accuracy increased with increase in ROIs indicating that finer resolution or detail in the FC revealed more uniqueness. Mean accuracy using the geodesic distance was consistently higher than the mean accuracy using Pearson dissimilarity. For several conditions (*resting-state*, *language*, *motor*), accuracy using geodesic distance on FCs obtained with 100 ROIs was greater than accuracy obtained using Pearson dissimilarity with 400 ROIs (Fig. C.5).

C.4 Computing geodesic distances for matrices without full rank

Computing the geodesic distance between two FC matrices Q_1 and Q_2 (Equation 4.3) requires Q_1 to be invertible, or equivalently, all the eigenvalues of Q_1 must

be strictly greater than zero. When FC matrices are based on n ROIs and n is larger than number of frames in the run, the *rank* of the resulting FC matrix is not full (i.e., $< n$), and some of its some eigenvalues are equal to 0. In practice, when the number of ROIs $n < (0.9 \times$ number of frames), we applied the procedure below to ensure full rankness.

To handle such cases, we adopted a simple approach here: we added the identity matrix I to both Q_1 and Q_2, causing the eigenvalues of the correlation matrices of interest to be increased by 1. Because all eigenvalues are then greater than 0, the matrices are invertible. In such cases, the geodesic distance, $d_G(Q_1 + I, Q_2 + I)$, serves as a proxy for the geodesic distance between the two matrices. Note that the scenario of low-rank FC matrices arises only for whole-cortex analysis, as for the subnetwork analyses, the number of ROIs in question was always greater than the number of frames in the run.

For reference, the procedure above was employed in the following cases: whole-cortex results for all tasks; whole-cortex *resting-state* results with lengths less than 400 TRs; and whole-cortex results involving trimmed data.

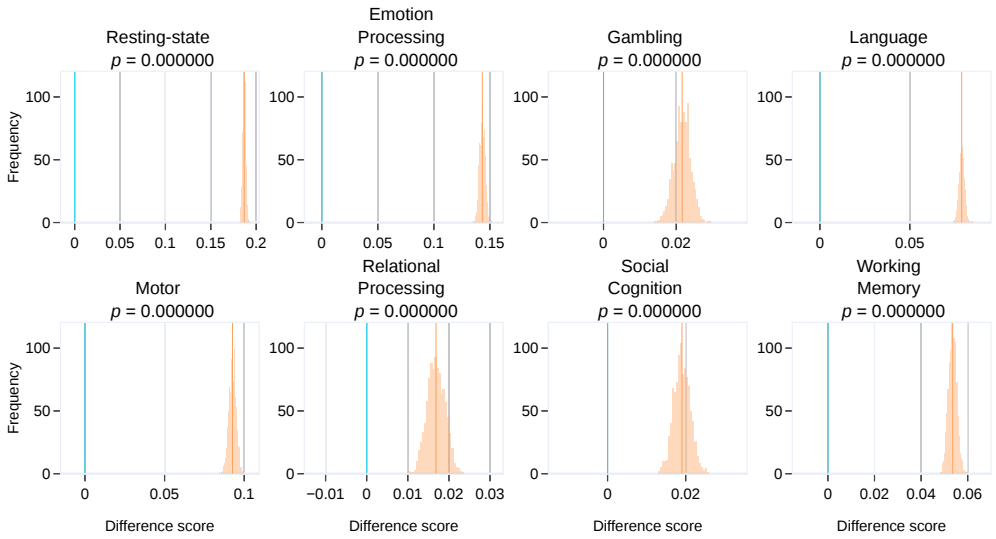

Figure C.6: *Whole-cortex FCs with full time course lengths:* Comparison of identification accuracy based on geodesic distance and Pearson dissimilarity for each condition. Identification was based on whole-cortex FCs. Here, full time course lengths were used (see Section 4.1.1). For each condition, the distributions shown in orange represent the difference between the mean participant identification accuracy using the geodesic distance and Pearson dissimilarity across the outer bootstrap iterations (see Section 4.1.7). The orange line indicates the mean of the difference distribution and the blue line indicates zero difference.

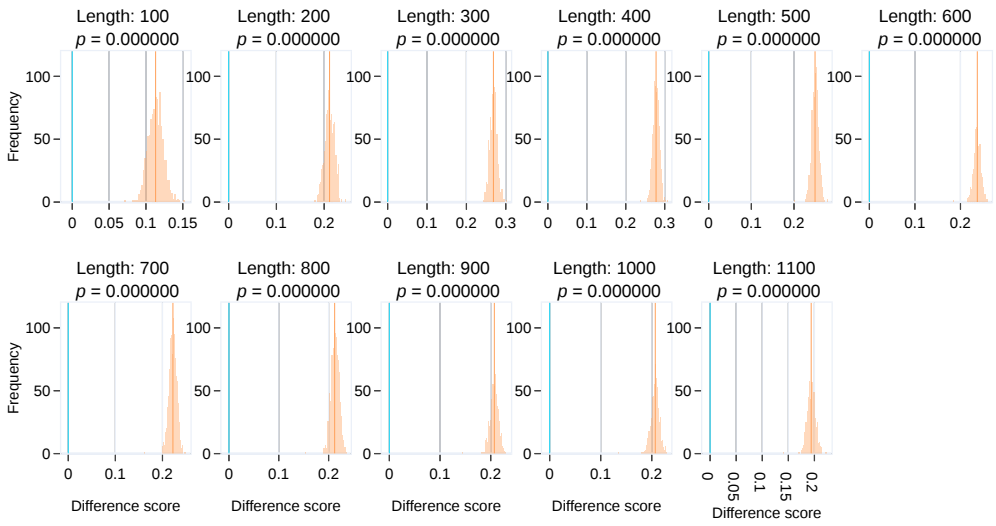

Figure C.7: *Identification accuracy and time course length:* Comparison of identification accuracy based on geodesic distance and Pearson dissimilarity for various time course lengths. Since *resting-state* data had the highest time course length, smaller segments of various lengths were extracted (see Section 4.1.7.1). Identification was based on whole-cortex FCs. For each segment length, the distributions shown in orange represent the difference between the mean participant identification accuracy using the geodesic distance and Pearson dissimilarity across the outer bootstrap iterations (see Section 4.1.7). The orange line indicates the mean of the difference distribution and the blue line indicates zero difference.

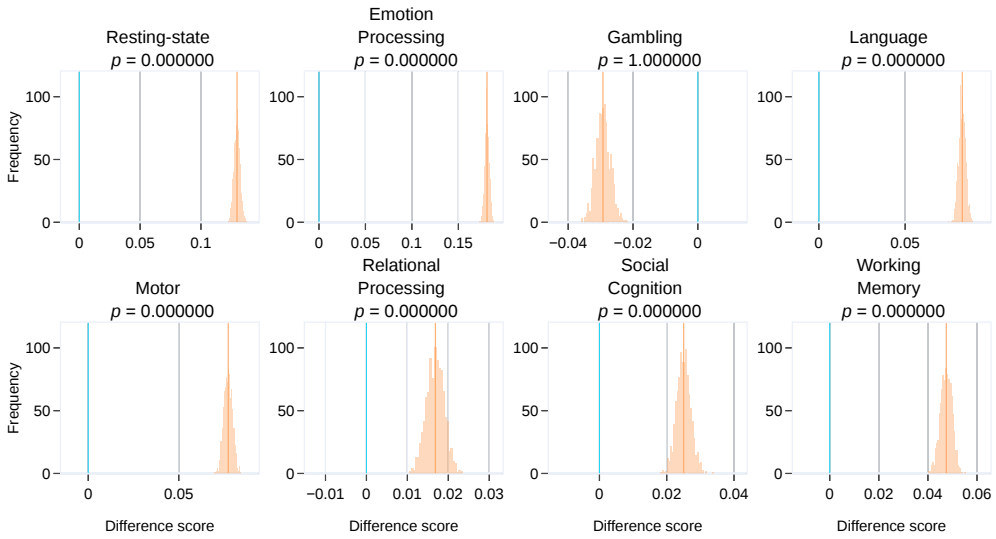

Figure C.8: *Whole-cortex FCs with trimmed time course lengths:* Comparison of identification accuracy based on geodesic distance and Pearson dissimilarity for each condition. Identification was based on whole-cortex FCs. Data for each condition were trimmed such that they all had the same time course length (of 138; see Section 4.2.5). For each condition, the distributions shown in orange represent the difference between the mean participant identification accuracy using the geodesic distance and Pearson dissimilarity across the outer bootstrap iterations (see Section 4.1.7). The orange line indicates the mean of the difference distribution and the blue line indicates zero difference.

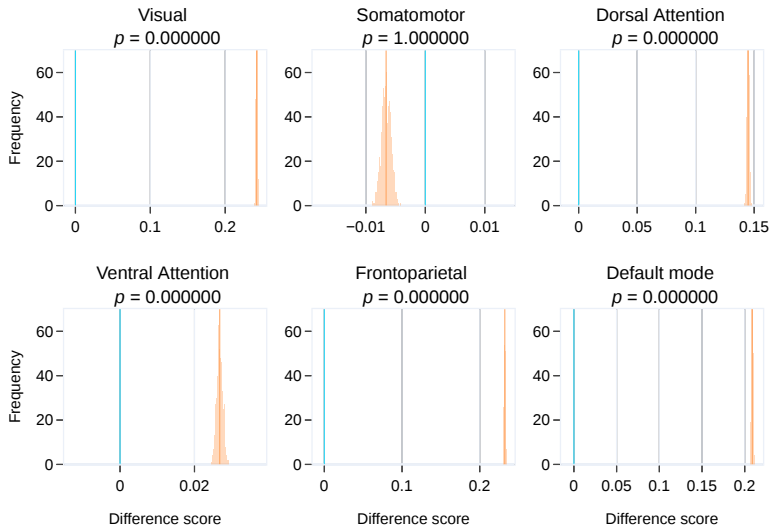

Figure C.9: *Subnetwork FCs with trimmed time course lengths:* Comparison of identification accuracy based on geodesic distance and Pearson dissimilarity for each subnetwork. Identification was based on subnetwork FCs. Data for each condition were trimmed such that they had the same time course length (of 138; see Section 4.2.5). For each subnetwork, difference scores were averaged across all conditions. The distributions shown in orange represent the difference between the mean participant identification accuracy using the geodesic distance and Pearson dissimilarity across the outer bootstrap iterations (see Section 4.1.7). The orange line indicates the mean of the difference distribution and the blue line indicates zero difference.

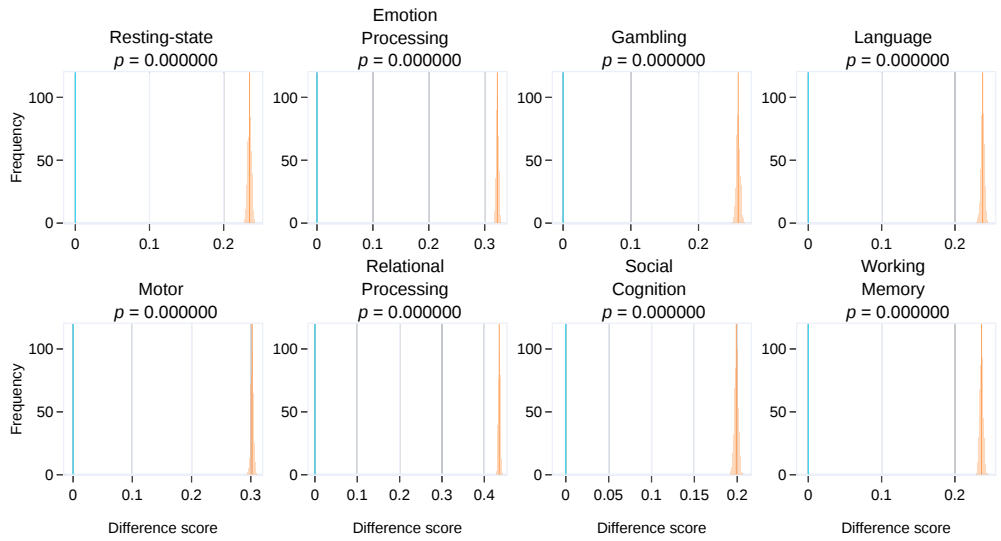

Figure C.10: *Subnetworks of the same size:* Comparison of identification accuracy using `dorsal attention` and `ventral attention` subnetwork FCs for each condition. The geodesic distance measure was used for identification. The two subnetworks were of identical size for the 300 ROIs parcellation (see Table 4.2. Data for each condition were trimmed such that they all had the same time course length (of 138; see Section 4.2.5). For each condition, the distributions shown in orange represent the difference between the mean participant identification accuracy based on the two subnetworks across the outer bootstrap iterations (see Section 4.1.7). The orange line indicates the mean of the difference distribution and the blue line indicates zero difference.

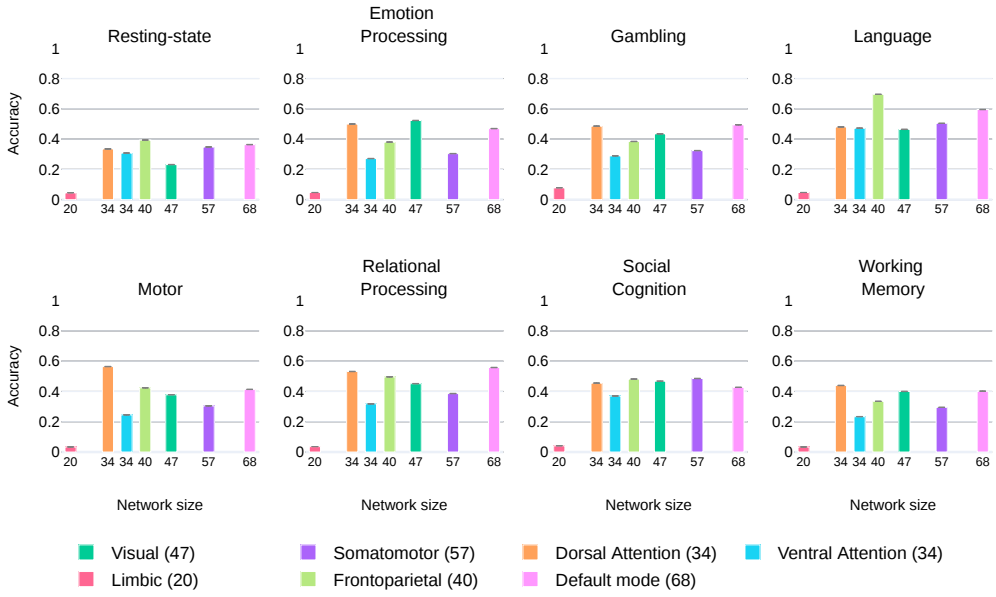

Figure C.11: Participant identification accuracy plotted against subnetwork size for each condition (Pearson dissimilarity). The size of the subnetwork (the number of ROIs) is also indicated in the inset. The error bars represent standard error of the mean across bootstrap iterations.

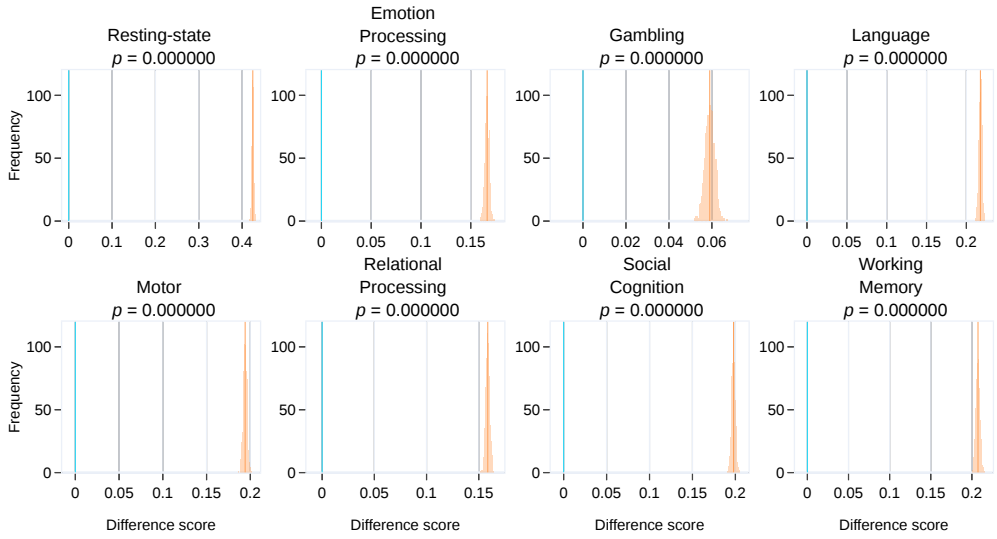

Figure C.12: *Combined subnetwork FCs with trimmed time course lengths:* Comparison of identification accuracy based on geodesic distance and Pearson dissimilarity for each condition. Identification was based on combined subnetwork FCs (see Section 4.2.7). Data for each condition were trimmed such that they all had the same time course length (of 138; see Section 4.2.5). For each condition, the distributions shown in orange represent the difference between the mean participant identification accuracy using the geodesic distance and Pearson dissimilarity across the outer bootstrap iterations (see Section 4.1.7). The orange line indicates the mean of the difference distribution and the blue line indicates zero difference.

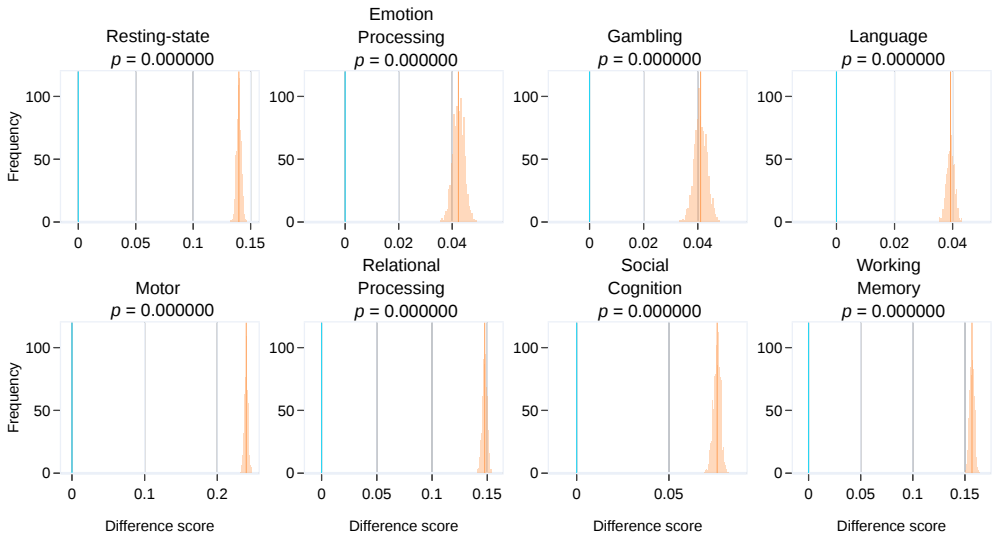

Figure C.13: *Combined subnetwork vs frontoparietal:* Comparison of identification accuracy based on combined subnetwork FCs (see Section 4.2.7) and `frontoparietal` subnetwork FCs (part of the combined subnetwork) for each condition. The geodesic distance measure was used for identification. Data for each condition were trimmed such that they all had the same time course length (of 138; see Section 4.2.5). For each condition, the distributions shown in orange represent the difference between the mean participant identification accuracy based on combined subnetwork FCs and `frontoparietal` FCs across the outer bootstrap iterations (see Section 4.1.7). The orange line indicates the mean of the difference distribution and the blue line indicates zero difference.

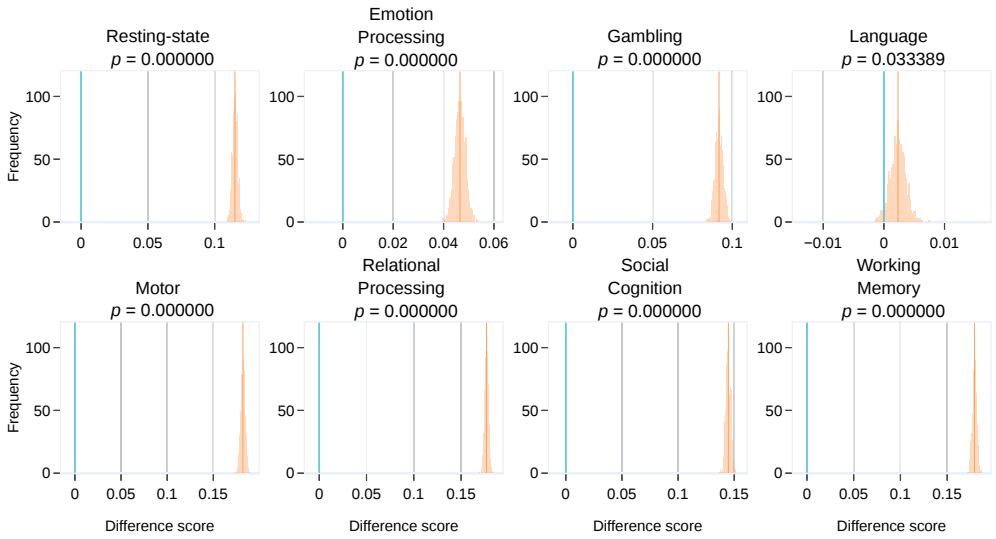

Figure C.14: *Combined subnetwork vs default mode:* Comparison of identification accuracy based on combined subnetwork FCs (see Section 4.2.7) and `default mode` subnetwork FCs (part of the combined subnetwork) for each condition. The geodesic distance measure was used for identification. Data for each condition were trimmed such that they all had the same time course length (of 138; see Section 4.2.5). For each condition, the distributions shown in orange represent the difference between the mean participant identification accuracy based on combined subnetwork FCs and `default mode` FCs across the outer bootstrap iterations (see Section 4.1.7). The orange line indicates the mean of the difference distribution and the blue line indicates zero difference.

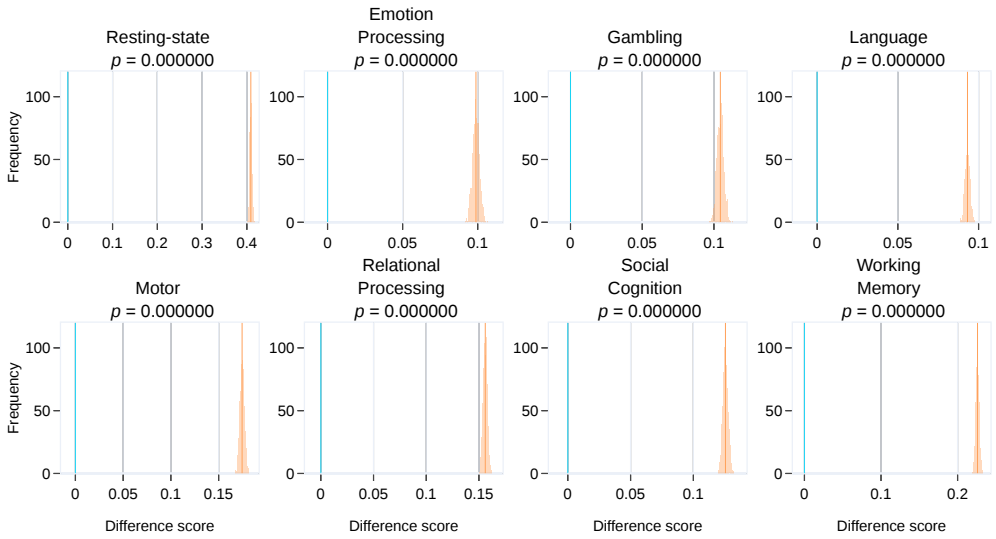

Figure C.15: *Combined subnetwork vs whole-cortex FCs:* Comparison of identification accuracy based on combined subnetwork FCs (see Section 4.2.7) and whole-cortex FCs for each condition. The geodesic distance measure was used for identification. Data for each condition were trimmed such that they all had the same time course length (of 138; see Section 4.2.5). For each condition, the distributions shown in orange represent the difference between the mean participant identification accuracy based on combined subnetwork FCs and whole-cortex FCs across the outer bootstrap iterations (see Section 4.1.7). The orange line indicates the mean of the difference distribution and the blue line indicates zero difference.

Bibliography

[1] P. P. Broca, "Loss of speech, chronic softening and partial destruction of the anterior left lobe of the brain," *Trans. revised*, 2003.

[2] L. Pessoa, "Understanding brain networks and brain organization," *Physics of Life Reviews*, vol. 11, no. 3, pp. 400–435, 2014.

[3] D. V. Buonomano and W. Maass, "State-dependent computations: spatiotemporal processing in cortical networks," *Nature Reviews Neuroscience*, vol. 10, no. 2, p. 113, 2009.

[4] J.-D. Haynes, "A primer on pattern-based approaches to fmri: principles, pitfalls, and perspectives," *Neuron*, vol. 87, no. 2, pp. 257–270, 2015.

[5] J. Mourao-Miranda, K. J. Friston, and M. Brammer, "Dynamic discrimination analysis: a spatial–temporal svm," *NeuroImage*, vol. 36, no. 1, pp. 88–99, 2007.

[6] R. A. Hutchinson, R. S. Niculescu, T. A. Keller, I. Rustandi, and T. M. Mitchell, "Modeling fmri data generated by overlapping cognitive processes with unknown onsets using hidden process models," *NeuroImage*, vol. 46, no. 1, pp. 87–104, 2009.

[7] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural Computation*, vol. 1, no. 2, pp. 270–280, 1989.

[8] B. A. Pearlmutter, "Learning state space trajectories in recurrent neural networks," *Neural Computation*, vol. 1, no. 2, pp. 263–269, 1989.

[9] B. G. Horne and C. L. Giles, "An experimental comparison of recurrent neural networks," in *Advances in Neural Information Processing Systems*, 1995, pp. 697–704.

[10] B. M. Yu, J. P. Cunningham, G. Santhanam, S. I. Ryu, K. V. Shenoy, and M. Sahani, "Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity," in *Advances in Neural Information*

Processing Systems 21, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds. Curran Associates, Inc., 2009, pp. 1881–1888.

[11] P. Gao, E. Trautmann, M. Y. Byron, G. Santhanam, S. Ryu, K. Shenoy, and S. Ganguli, "A theory of multineuronal dimensionality, dynamics and measurement," *bioRxiv*, p. 214262, 2017.

[12] N. K. Logothetis, "The underpinnings of the bold functional magnetic resonance imaging signal," *Journal of Neuroscience*, vol. 23, no. 10, pp. 3963–3971, 2003.

[13] Y. Ma, M. A. Shaik, M. G. Kozberg, S. H. Kim, J. P. Portes, D. Timerman, and E. M. Hillman, "Resting-state hemodynamics are spatiotemporally coupled to synchronized and symmetric neural activity in excitatory neurons," *Proceedings of the National Academy of Sciences*, vol. 113, no. 52, pp. E8463–E8471, 2016.

[14] S. M. Smith, T. E. Nichols, D. Vidaurre, A. M. Winkler, T. E. Behrens, M. F. Glasser, K. Ugurbil, D. M. Barch, D. C. Van Essen, and K. L. Miller, "A positive-negative mode of population covariation links brain connectivity, demographics and behavior," *Nature neuroscience*, vol. 18, no. 11, pp. 1565–1567, 2015.

[15] A. T. Drysdale, L. Grosenick, J. Downar, K. Dunlop, F. Mansouri, Y. Meng, R. N. Fetcho, B. Zebley, D. J. Oathes, A. Etkin *et al.*, "Resting-state connectivity biomarkers define neurophysiological subtypes of depression," *Nature medicine*, vol. 23, no. 1, pp. 28–38, 2017.

[16] C. H. Xia, Z. Ma, R. Ciric, S. Gu, R. F. Betzel, A. N. Kaczkurkin, M. E. Calkins, P. A. Cook, A. G. de la Garza, S. N. Vandekar *et al.*, "Linked dimensions of psychopathology and connectivity in functional brain networks," *Nature communications*, vol. 9, no. 1, pp. 1–14, 2018.

[17] J. M. Zacks, T. S. Braver, M. A. Sheridan, D. I. Donaldson, A. Z. Snyder, J. M. Ollinger, R. L. Buckner, and M. E. Raichle, "Human brain activity time-locked to perceptual event boundaries," *Nature Neuroscience*, vol. 4, no. 6, p. 651, 2001.

[18] M. Najafi, J. Kinnison, and L. Pessoa, "Dynamics of intersubject brain networks during anxious anticipation," *Frontiers in Human Neuroscience*, vol. 11, p. 552, 2017.

[19] E. C. Ferstl, M. Rinck, and D. Y. v. Cramon, "Emotional and temporal aspects of situation model processing during text comprehension: An event-related fmri study," *Journal of Cognitive Neuroscience*, vol. 17, no. 5, pp. 724–739, 2005.

[20] U. Hasson, Y. Nir, I. Levy, G. Fuhrmann, and R. Malach, "Intersubject synchronization of cortical activity during natural vision," *Science*, vol. 303, no. 5664, pp. 1634–1640, 2004.

[21] M. A. Bertolero and D. S. Bassett, "On the nature of explanations offered by network science: A perspective from and for practicing neuroscientists," *Topics in Cognitive Science*, 2020.

[22] C. Zednik, "Models and mechanisms in network neuroscience," *Philosophical Psychology*, vol. 32, no. 1, pp. 23–51, 2019.

[23] M. P. Van Den Heuvel and A. Fornito, "Brain networks in schizophrenia," *Neuropsychology review*, vol. 24, no. 1, pp. 32–48, 2014.

[24] R. H. Kaiser, J. R. Andrews-Hanna, T. D. Wager, and D. A. Pizzagalli, "Large-scale network dysfunction in major depressive disorder: a meta-analysis of resting-state functional connectivity," *JAMA psychiatry*, vol. 72, no. 6, pp. 603–611, 2015.

[25] M. D. Rosenberg, E. S. Finn, D. Scheinost, X. Papademetris, X. Shen, R. T. Constable, and M. M. Chun, "A neuromarker of sustained attention from whole-brain functional connectivity," *Nature neuroscience*, vol. 19, no. 1, pp. 165–171, 2016.

[26] C. Olah, "Understanding lstm networks, 2015," *URL http://colah. github. io/posts/2015-08-Understanding-LSTMs*, 2015.

[27] M. Venkatesh, J. Jaja, and L. Pessoa, "Brain dynamics and temporal trajectories during task and naturalistic processing," *NeuroImage*, vol. 186, pp. 410–423, 2019.

[28] S. Sonkusare, M. Breakspear, and C. Guo, "Naturalistic stimuli in neuroscience: Critically acclaimed," *Trends in Cognitive Sciences*, vol. 23, no. 8, pp. 699–714, 2019.

[29] P. Bartolomeo, T. S. Malkinson, and S. De Vito, "Botallo's error, or the quandaries of the universality assumption," *Cortex*, vol. 86, pp. 176–185, 2017.

[30] R. E. Beaty, Y. N. Kenett, A. P. Christensen, M. D. Rosenberg, M. Benedek, Q. Chen, A. Fink, J. Qiu, T. R. Kwapil, M. J. Kane *et al.*, "Robust prediction of individual creative ability from brain functional connectivity," *Proceedings of the National Academy of Sciences*, vol. 115, no. 5, pp. 1087–1092, 2018.

[31] J. Dubois, P. Galdi, L. K. Paul, and R. Adolphs, "A distributed brain network predicts general intelligence from resting-state human neuroimaging data," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 373, no. 1756, p. 20170284, 2018.

[32] J. Dubois, P. Galdi, Y. Han, L. K. Paul, and R. Adolphs, "Resting-state functional brain connectivity best predicts the personality dimension of openness to experience," *Personality Neuroscience*, vol. 1, 2018.

[33] W.-T. Hsu, M. D. Rosenberg, D. Scheinost, R. T. Constable, and M. M. Chun, "Resting-state functional connectivity predicts neuroticism and extraversion in novel individuals," *Social Cognitive and Affective Neuroscience*, vol. 13, no. 2, pp. 224–232, 2018.

[34] R. Jiang, V. D. Calhoun, N. Zuo, D. Lin, J. Li, L. Fan, S. Qi, H. Sun, Z. Fu, M. Song *et al.*, "Connectome-based individualized prediction of temperament trait scores," *NeuroImage*, vol. 183, pp. 366–374, 2018.

[35] M. Lukoševicius, "Echo state networks with trained feedbacks," *Networks*, no. 4, 2007.

[36] M. Lukoeviius and H. Jaeger, "Reservoir computing approaches to recurrent neural network training," *Computer Science Review*, vol. 3, no. 3, pp. 127 – 149, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1574013709000173

[37] F. Wyffels, B. Schrauwen, and D. Stroobandt, "Stable output feedback in reservoir computing using ridge regression," in *International conference on artificial neural networks*. Springer, 2008, pp. 808–817.

[38] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[39] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[40] E. A. Allen, E. Damaraju, S. M. Plis, E. B. Erhardt, T. Eichele, and V. D. Calhoun, "Tracking Whole-Brain Connectivity Dynamics in the Resting State," *Cerebral Cortex*, vol. 24, pp. 663–676, 2014.

[41] E. S. Finn, X. Shen, D. Scheinost, M. D. Rosenberg, J. Huang, M. M. Chun, X. Papademetris, and R. T. Constable, "Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity," *Nature Neuroscience*, vol. 18, no. 11, p. 1664, 2015.

[42] M. Venkatesh, J. Jaja, and L. Pessoa, "Comparing functional connectivity matrices: A geometry-aware approach applied to participant identification," *NeuroImage*, vol. 207, p. 116398, 2020.

[43] S. A. Huettel, A. W. Song, G. McCarthy *et al.*, *Functional magnetic resonance imaging*. Sinauer Associates Sunderland, 2004, vol. 1.

[44] J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini, "Distributed and overlapping representations of faces and objects in ventral temporal cortex," *Science*, vol. 293, no. 5539, pp. 2425–2430, 2001.

[45] Y. Kamitani and F. Tong, "Decoding the visual and subjective contents of the human brain," *Nature Neuroscience*, vol. 8, no. 5, p. 679, 2005.

[46] J.-D. Haynes and G. Rees, "Neuroimaging: decoding mental states from brain activity in humans," *Nature Reviews Neuroscience*, vol. 7, no. 7, p. 523, 2006.

[47] D. D. Cox and R. L. Savoy, "Functional magnetic resonance imaging (fmri)brain reading: detecting and classifying distributed patterns of fmri activity in human visual cortex," *NeuroImage*, vol. 19, no. 2, pp. 261–270, 2003.

[48] N. Kriegeskorte, R. Goebel, and P. Bandettini, "Information-based functional brain mapping," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 10, pp. 3863–3868, 2006.

[49] A. Nestor, D. C. Plaut, and M. Behrmann, "Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis," *Proceedings of the National Academy of Sciences*, vol. 108, no. 24, pp. 9998–10 003, 2011.

[50] F. Janoos, R. Machiraju, S. Singh, and I. Á. Morocz, "Spatio-temporal models of mental processes from fmri," *NeuroImage*, vol. 57, no. 2, pp. 362–377, 2011.

[51] C. Chu, J. Mourão-Miranda, Y.-C. Chiu, N. Kriegeskorte, G. Tan, and J. Ashburner, "Utilizing temporal information in fmri decoding: classifier using kernel regression methods," *NeuroImage*, vol. 58, no. 2, pp. 560–571, 2011.

[52] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International Conference on Machine Learning*, 2013, pp. 1310–1318.

[53] J. Martens and I. Sutskever, "Learning recurrent neural networks with hessian-free optimization," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. Citeseer, 2011, pp. 1033–1040.

[54] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 6645–6649.

[55] W. Maass, T. Natschläger, and H. Markram, "Real-time computing without stable states: A new framework for neural computation based on perturbations," *Neural Computation*, vol. 14, no. 11, pp. 2531–2560, 2002.

[56] H. Jaeger, "The echo state approach to analysing and training recurrent neural networks-with an erratum note," *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report*, vol. 148, no. 34, p. 13, 2001.

[57] H. Jaeger and H. Haas, "Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication," *Science*, vol. 304, no. 5667, pp. 78–80, 2004.

[58] J. J. Steil, "Backpropagation-decorrelation: online recurrent learning with o (n) complexity," in *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, vol. 2. IEEE, 2004, pp. 843–848.

[59] D. Sussillo and L. F. Abbott, "Generating coherent patterns of activity from chaotic neural networks," *Neuron*, vol. 63, no. 4, pp. 544–557, 2009.

[60] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE Transactions on Electronic Computers*, no. 3, pp. 326–334, 1965.

[61] Z. Lu, J. Pathak, B. Hunt, M. Girvan, R. Brockett, and E. Ott, "Reservoir observers: Model-free inference of unmeasured variables in chaotic systems," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 27, no. 4, p. 041102, 2017.

[62] M. D. Skowronski and J. G. Harris, "Automatic speech recognition using a predictive echo state network classifier," *Neural Networks*, vol. 20, no. 3, pp. 414–423, 2007.

[63] F. Triefenbach, A. Jalalvand, B. Schrauwen, and J.-P. Martens, "Phoneme recognition with large hierarchical reservoirs," in *Advances in Neural Information Processing Systems*, 2010, pp. 2307–2315.

[64] K. Vandoorne, W. Dierckx, B. Schrauwen, D. Verstraeten, R. Baets, P. Bienstman, and J. Van Campenhout, "Toward optical signal processing using photonic reservoir computing," *Optics Express*, vol. 16, no. 15, pp. 11 182–11 192, 2008.

[65] D. M. Barch, G. C. Burgess, M. P. Harms, S. E. Petersen, B. L. Schlaggar, M. Corbetta, M. F. Glasser, S. Curtiss, S. Dixit, C. Feldt *et al.*, "Function in the human connectome: task-fmri and individual differences in behavior," *NeuroImage*, vol. 80, pp. 169–189, 2013.

[66] D. A. Feinberg, S. Moeller, S. M. Smith, E. Auerbach, S. Ramanna, M. F. Glasser, K. L. Miller, K. Ugurbil, and E. Yacoub, "Multiplexed echo planar imaging for sub-second whole brain fmri and fast diffusion imaging," *PloS One*, vol. 5, no. 12, p. e15710, 2010.

[67] R. W. Cox, "Afni: software for analysis and visualization of functional magnetic resonance neuroimages," *Computers and Biomedical Research*, vol. 29, no. 3, pp. 162–173, 1996.

[68] J. E. Iglesias, C.-Y. Liu, P. M. Thompson, and Z. Tu, "Robust brain extraction across datasets and comparison with publicly available methods," *IEEE Transactions on Medical Imaging*, vol. 30, no. 9, pp. 1617–1634, 2011.

[69] D. N. Greve and B. Fischl, "Accurate and robust brain image alignment using boundary-based registration," *Neuroimage*, vol. 48, no. 1, pp. 63–72, 2009.

[70] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee, "A reproducible evaluation of ants similarity metric performance in brain image registration," *NeuroImage*, vol. 54, no. 3, pp. 2033–2044, 2011.

[71] J. F. Smith, J. Hur, C. M. Kaplan, and A. J. Shackman, "The impact of spatial normalization for functional magnetic resonance imaging data analyses revisited," *bioRxiv*, p. 272302, 2018.

[72] M. F. Glasser, T. S. Coalson, E. C. Robinson, C. D. Hacker, J. Harwell, E. Yacoub, K. Ugurbil, J. Andersson, C. F. Beckmann, M. Jenkinson *et al.*, "A multi-modal parcellation of human cerebral cortex," *Nature*, vol. 536, no. 7615, pp. 171–178, 2016.

[73] B. M. Nacewicz, L. Angelos, K. M. Dalton, R. Fischer, M. J. Anderle, A. L. Alexander, and R. J. Davidson, "Reliable non-invasive measurement of human neurochemistry using proton spectroscopy with an anatomically defined amygdala-specific voxel," *NeuroImage*, vol. 59, no. 3, pp. 2548–2559, 2012.

[74] M. Lukoševičius, "A practical guide to applying echo state networks," in *Neural networks: Tricks of the trade*. Springer, 2012, pp. 659–686.

[75] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[76] B. Scholkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[77] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[78] R. G. Brereton and G. R. Lloyd, "Partial least squares discriminant analysis: taking the magic away," *Journal of Chemometrics*, vol. 28, no. 4, pp. 213–225, 2014.

[79] M. Ojala and G. C. Garriga, "Permutation tests for studying classifier performance," *Journal of Machine Learning Research*, vol. 11, no. Jun, pp. 1833–1863, 2010.

[80] E. Bullmore, J. Fadili, V. Maxim, L. Şendur, B. Whitcher, J. Suckling, M. Brammer, and M. Breakspear, "Wavelets and functional magnetic resonance imaging of the human brain," *NeuroImage*, vol. 23, pp. S234–S249, 2004.

[81] M. Schurz, J. Radua, M. Aichhorn, F. Richlan, and J. Perner, "Fractionating theory of mind: a meta-analysis of functional brain imaging studies," *Neuroscience and Biobehavioral Reviews*, vol. 42, pp. 9–34, 2014.

[82] J. M. Shine, M. Breakspear, P. Bell, K. E. Martens, R. Shine, O. Koyejo, O. Sporns, and R. Poldrack, "The low dimensional dynamic and integrative core of cognition in the human brain," *bioRxiv*, p. 266635, 2018.

[83] P. Enel, E. Procyk, R. Quilodran, and P. F. Dominey, "Reservoir computing properties of neural dynamics in prefrontal cortex," *PLoS Computational Biology*, vol. 12, no. 6, p. e1004967, 2016.

[84] J. A. Bartz, J. Zaki, N. Bolger, and K. N. Ochsner, "Social effects of oxytocin in humans: context and person matter," *Trends in Cognitive Sciences*, vol. 15, no. 7, pp. 301–309, 2011.

[85] Y. Lerner, C. J. Honey, L. J. Silbert, and U. Hasson, "Topographic mapping of a hierarchy of temporal receptive windows using a narrated story," *Journal of Neuroscience*, vol. 31, no. 8, pp. 2906–2915, 2011.

[86] G. Barbera, B. Liang, L. Zhang, C. R. Gerfen, E. Culurciello, R. Chen, Y. Li, and D.-T. Lin, "Spatially compact neural clusters in the dorsal striatum encode locomotion relevant information," *Neuron*, vol. 92, no. 1, pp. 202–213, 2016.

[87] M. Corbetta and G. L. Shulman, "Control of goal-directed and stimulus-driven attention in the brain," *Nature Reviews Neuroscience*, vol. 3, no. 3, p. 201, 2002.

[88] L. Pessoa and L. G. Ungerleider, "Top-down mechanisms for working memory and attentional processes." *APA*, 2004.

[89] S. J. Carrington and A. J. Bailey, "Are there theory of mind regions in the brain? a review of the neuroimaging literature," *Human Brain Mapping*, vol. 30, no. 8, pp. 2313–2335, 2009.

[90] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, W.-M. H. Consortium *et al.*, "The wu-minn human connectome project: an overview," *NeuroImage*, vol. 80, pp. 62–79, 2013.

[91] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni *et al.*, "The minimal preprocessing pipelines for the human connectome project," *NeuroImage*, vol. 80, pp. 105–124, 2013.

[92] A. T. Vu, K. Jamison, M. F. Glasser, S. M. Smith, T. Coalson, S. Moeller, E. J. Auerbach, K. Uğurbil, and E. Yacoub, "Tradeoffs in pushing the spatial resolution of fmri for the 7t human connectome project," *NeuroImage*, vol. 154, pp. 23–32, 2017.

[93] A. Schaefer, R. Kong, E. M. Gordon, T. O. Laumann, X.-N. Zuo, A. J. Holmes, S. B. Eickhoff, and B. T. Yeo, "Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri," *Cerebral Cortex*, vol. 28, no. 9, pp. 3095–3114, 2018.

[94] B. T. T. Yeo, F. M. Krienen, J. Sepulcre, M. R. Sabuncu, D. Lashkari, M. Hollinshead, J. L. Roffman, J. W. Smoller, L. Zllei, J. R. Polimeni, B. Fischl, H. Liu, and R. L. Buckner, "The organization of the human cerebral cortex estimated by intrinsic functional connectivity," *Journal of Neurophysiology*, vol. 106, no. 3, pp. 1125–1165, 2011.

[95] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[96] F. A. Gers, J. A. Schmidhuber, and F. A. Cummins, "Learning to forget: Continual prediction with lstm," *Neural Computation*, vol. 12, no. 10, p. 24512471, Oct. 2000.

[97] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[98] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, 2019, pp. 8024–8035.

[99] G. Alain and Y. Bengio, "Understanding intermediate layers using linear classifier probes," *arXiv preprint arXiv:1610.01644*, 2016.

[100] G. Montavon, M. L. Braun, and K.-R. Müller, "Kernel analysis of deep networks," *Journal of Machine Learning Research*, vol. 12, no. 78, pp. 2563–2581, 2011.

[101] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[102] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[103] E. S. Finn, E. Glerean, A. Y. Khojandi, D. Nielson, P. J. Molfese, D. A. Handwerker, and P. A. Bandettini, "Idiosynchrony: From shared responses to individual differences during naturalistic neuroimaging," *NeuroImage*, vol. 215, p. 116828, 2020.

[104] X. Shen, E. S. Finn, D. Scheinost, M. D. Rosenberg, M. M. Chun, X. Papademetris, and R. T. Constable, "Using connectome-based predictive modeling to predict individual behavior from brain connectivity," *Nature Protocols*, vol. 12, no. 3, p. 506, 2017.

[105] O. Sporns, *Networks of the Brain*. MIT press, 2010.

[106] E. C. Hansen, D. Battaglia, A. Spiegler, G. Deco, and V. K. Jirsa, "Functional connectivity dynamics: Modeling the switching behavior of the resting state," *NeuroImage*, vol. 105, pp. 525–535, 2015.

[107] D. H. Schultz and M. W. Cole, "Higher intelligence is associated with less task-related brain network reconfiguration," *Journal of Neuroscience*, vol. 36, no. 33, pp. 8551–8561, 2016.

[108] E. S. Finn, D. Scheinost, D. M. Finn, X. Shen, X. Papademetris, and R. T. Constable, "Can brain state be manipulated to emphasize individual differences in functional connectivity?" *NeuroImage*, vol. 160, pp. 140–151, 10 2017.

[109] E. Amico and J. Goñi, "The quest for identifiability in human functional connectomes," *Scientific Reports*, vol. 8, no. 1, p. 8254, 2018.

[110] V. Ponsoda, K. Martínez, J. A. Pineda-Pardo, F. J. Abad, J. Olea, F. J. Román, A. K. Barbey, and R. Colom, "Structural brain connectivity and cognitive ability differences: A multivariate distance matrix regression analysis," *Human Brain Mapping*, vol. 38, no. 2, pp. 803–816, 2017.

[111] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *International Conference on Database Theory*. Springer, 2001, pp. 420–434.

[112] D. C. V. Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, and f. t. W.-M. H. Consortium, "The WU-Minn Human Connectome Project: An overview," *NeuroImage*, vol. 80, pp. 62–79, 2013.

[113] D. M. Barch, G. C. Burgess, M. P. Harms, S. E. Petersen, B. L. Schlaggar, M. Corbetta, M. F. Glasser, S. Curtiss, S. Dixit, C. Feldt, D. Nolan, E. Bryant, T. Hartley, O. Footer, J. M. Bjork, R. Poldrack, S. Smith, H. Johansen-Berg, A. Z. Snyder, D. C. V. Essen, and f. t. W.-M. H. Consortium, "Function in the human connectome: Task-fMRI and individual differences in behavior," *NeuroImage*, vol. 80, pp. 169–189, 2013.

[114] M. S. Cohen, "Parametric analysis of fmri data using linear systems methods," *NeuroImage*, vol. 6, no. 2, pp. 93–103, 1997.

[115] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, D. C. V. Essen, M. Jenkinson, and f. t. W.-M. H. Consortium, "The minimal preprocessing pipelines for the Human Connectome Project," *NeuroImage*, vol. 80, pp. 105–124, 2013.

[116] S. M. Smith, C. F. Beckmann, J. Andersson, E. J. Auerbach, J. Bijsterbosch, G. Douaud, E. Duff, D. A. Feinberg, L. Griffanti, M. P. Harms, M. Kelly, T. Laumann, K. L. Miller, S. Moeller, S. Petersen, J. Power, G. Salimi-Khorshidi, A. Z. Snyder, A. T. Vu, M. W. Woolrich, J. Xu, E. Yacoub, K. Uurbil, D. C. V. Essen, M. F. Glasser, and W.-M. H. Consortium, "Resting-state fMRI in the Human Connectome Project," *NeuroImage*, vol. 80, pp. 144–168, 2013.

[117] R. N. Boubela, K. Kalcher, W. Huf, C. Kronnerwetter, P. Filzmoser, and E. Moser, "Beyond noise: using temporal ica to extract meaningful information from high-frequency fmri signal fluctuations during rest," *Frontiers in Human Neuroscience*, vol. 7, p. 168, 2013.

[118] R. Kong, J. Li, C. Orban, M. R. Sabuncu, H. Liu, A. Schaefer, N. Sun, X.-N. Zuo, A. J. Holmes, S. B. Eickhoff *et al.*, "Spatial topography of individual-specific cortical networks predicts human cognition, personality, and emotion," *Cerebral Cortex*, vol. 29, no. 6, pp. 2533–2551, 2018.

[119] J. Li, R. Kong, R. Liegeois, C. Orban, Y. Tan, N. Sun, A. J. Holmes, M. R. Sabuncu, T. Ge, and B. T. Yeo, "Global signal regression strengthens association between resting-state functional connectivity and behavior," *NeuroImage*, vol. 196, pp. 126–141, 2019.

[120] A. Schaefer, R. Kong, E. M. Gordon, T. O. Laumann, X.-N. Zuo, A. J. Holmes, S. B. Eickhoff, and B. T. Yeo, "Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri," *Cerebral Cortex*, vol. 28, no. 9, pp. 3095–3114, 2017.

[121] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge university press, 2004.

[122] X. Pennec, P. Fillard, and N. Ayache, "A riemannian framework for tensor computing," *International Journal of Computer Vision*, vol. 66, no. 1, pp. 41–66, 2006.

[123] W. Förstner and B. Moonen, "A metric for covariance matrices," in *Geodesy-The Challenge of the 3rd Millennium*. Springer, 2003, pp. 299–309.

[124] S. Van Dongen and A. J. Enright, "Metric distances derived from cosine similarity and pearson and spearman correlations," *arXiv preprint arXiv:1208.3145*, 2012.

[125] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, 1964.

[126] MATLAB, *version 9.4.0.813654 (R2018a)*. Natick, Massachusetts: The MathWorks Inc., 2018.

[127] V. Amrhein, S. Greenland, and B. McShane, "Scientists rise up against statistical significance," 2019.

[128] B. B. McShane, D. Gal, A. Gelman, C. Robert, and J. L. Tackett, "Abandon Statistical Significance," *The American Statistician*, vol. 73, pp. 235–245, 2019.

[129] M. Venkatesh, "Online figures," https://github.com/makto-toruk/FC_geodesic, 2019.

[130] N. Leonardi and D. V. D. Ville, "On spurious and real fluctuations of dynamic functional connectivity during rest," *NeuroImage*, vol. 104, pp. 430–436, 2015.

[131] A. Zalesky and M. Breakspear, "Towards a statistical test for functional connectivity dynamics," *NeuroImage*, vol. 114, pp. 466–470, 2015.

[132] H. Xie, V. D. Calhoun, J. Gonzalez-Castillo, E. Damaraju, R. Miller, P. A. Bandettini, and S. Mitra, "Whole-brain connectivity dynamics reflect both task-specific and individual-specific modulation: A multitask study," *NeuroImage*, vol. 180, 2018.

[133] M. L. Elliott, A. R. Knodt, M. Cooke, M. J. Kim, T. R. Melzer, R. Keenan, D. Ireland, S. Ramrakha, R. Poulton, A. Caspi *et al.*, "General functional connectivity: shared features of resting-state and task fmri drive reliable and heritable individual differences in functional brain networks," *NeuroImage*, vol. 189, pp. 516–532, 2019.

[134] J. Gonzalez-Castillo, C. W. Hoy, D. A. Handwerker, M. E. Robinson, L. C. Buchanan, Z. S. Saad, and P. A. Bandettini, "Tracking ongoing cognition in individuals using brief, whole-brain functional connectivity patterns," *Proceedings of the National Academy of Sciences*, vol. 112, pp. 8762–8767, 2015.

[135] V. Estivill-Castro, "Why so many clustering algorithms: a position paper," *ACM SIGKDD Explorations Newsletter*, vol. 4, pp. 65–75, 2002.

[136] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*, 2002, pp. 849–856.

[137] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[138] K. Murphy and M. D. Fox, "Towards a consensus regarding global signal regression for resting state functional connectivity mri," *NeuroImage*, vol. 154, pp. 169–173, 2017.

[139] Z. S. Saad, S. J. Gotts, K. Murphy, G. Chen, H. J. Jo, A. Martin, and R. W. Cox, "Trouble at rest: how correlation patterns and group differences become distorted after global signal regression," *Brain Connectivity*, vol. 2, no. 1, pp. 25–32, 2012.